Variational Methods for Latent Variable Problems

Outline

- Latent variable problems
  - Examples
  - Why we need to integrate out latent variables (the Neyman-Scott paradox)
  - Integrating out the latent variable
- Bayesian and frequentist statistics
  - The difference between then
  - Integrating out latent variables in the two approaches
  - MAP and MLE estimates
- The EM algorithm
  - ...resolves the Neyman-Scott paradox
- Variational Bayes
  - KL divergence
  - Some intuition
- A bad banana example
- Conclusion & questions

Latent Variable Problems

Examples

Microcredit effectiveness

Randomized controlled trials were run in seven different countries to measure the effect of access to microcredit on business profits.

In each country, thousands of businesses were observed. These businesses share common, unobserved attributes of their particular country. We wish to infer the overall average effectiveness of microcredit.

Understanding the impact of microcredit expansions: A Bayesian hierarchical analysis of 7 randomised experiments. Meager, 2015

Examples

Topics in the New York Times

We imagine each article in the New York Times is about small number of topics. A topic is characterized by a distribution of words.

The words in a given article share a common, unobserved attribute: the topics of the article in which they appear. We wish to infer the identities of topics that appear in the NYT and their characteristic word distributions.

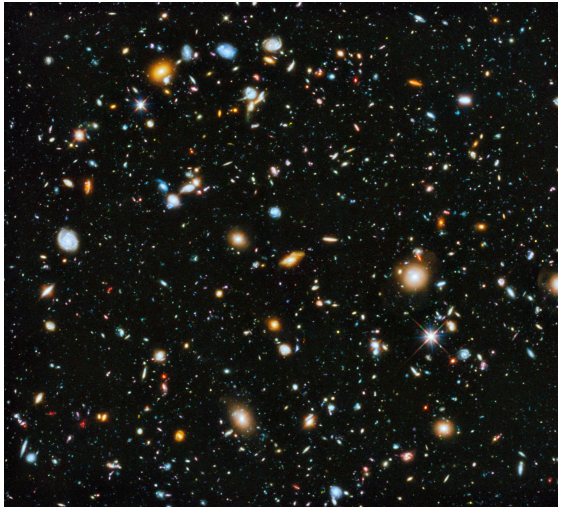
Latent Dirichlet Allocation. Blei, Ng, Jordan, 2003

Examples

Cosmology forward modeling

Given a set of initial states of the universe and a forward model characterized by a set of cosmological parameters, we can run a simulation and estimate the distribution of astronomical observations.

Observations from real-life telescopes share a common, unobserved attribute: the initial state of the universe. We want to infer the cosmological parameters.



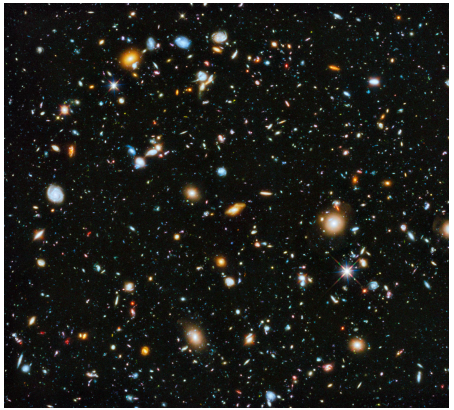
Towards optimal extraction of cosmological information from nonlinear data.
Seljak, Aslanyan, Feng, Modi (prepared for submission)

Examples

θ = Unobserved, relatively low-dimensional shared parameters

$z = (z_1, z_2, ..., z_n)$ = Unobserved, relatively high-dimensional set of "latent" parameters or data

X = Observed data



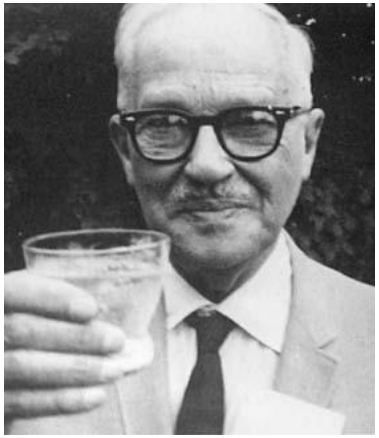
The Neyman-Scott "Paradox"

The Neyman-Scott paradox.

For $n = 1, ..., N$

$$X_{1n} \sim \mathcal{N}(z_n, \theta)$$

$$X_{2n} \sim \mathcal{N}(z_n, \theta)$$



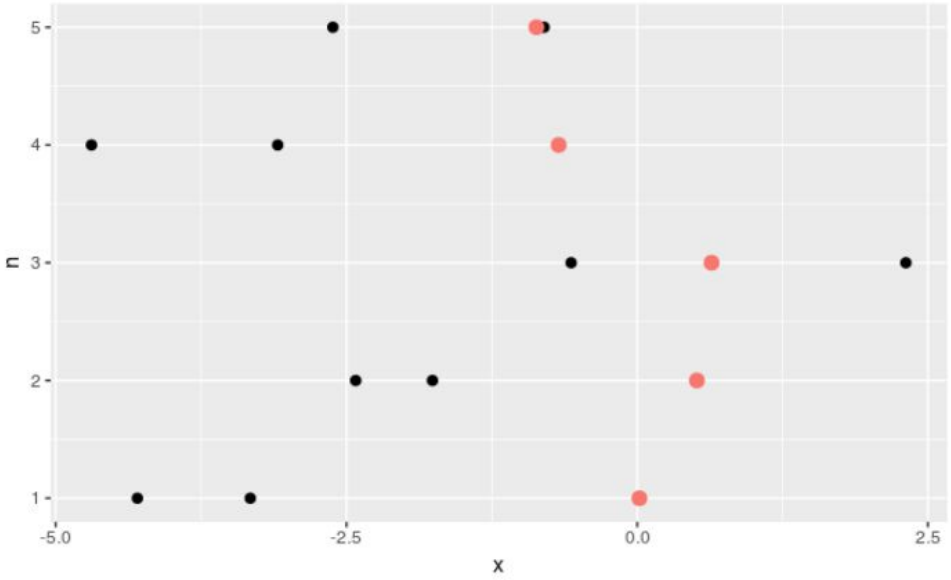
The Neyman-Scott "Paradox"

For $n = 1, ..., N$

$$X_{1n} \sim \mathcal{N}(z_n, \theta)$$

$$X_{2n} \sim \mathcal{N}(z_n, \theta)$$

We will investigate the "joint maximum likelihood estimator".



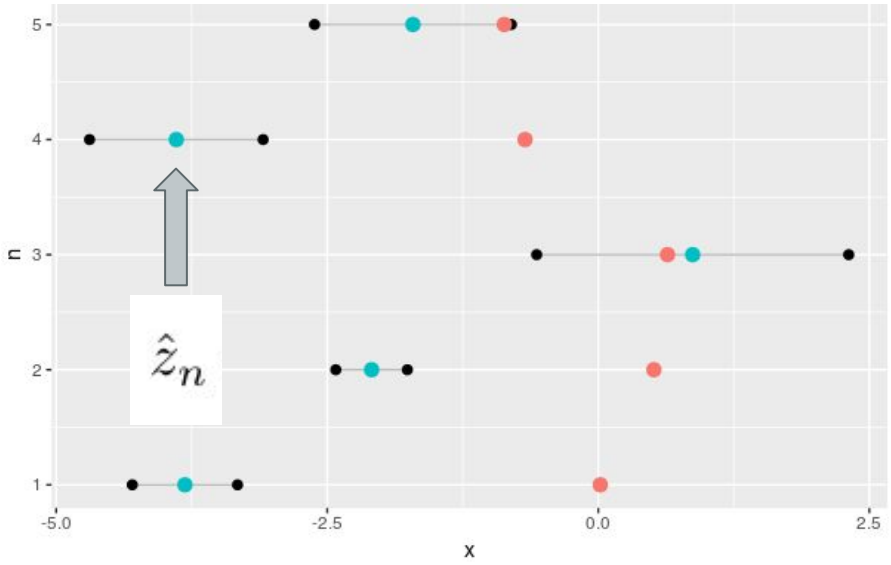
The Neyman-Scott "Paradox"

For $n = 1, ..., N$

$$X_{1n} \sim \mathcal{N}(z_n, \theta)$$
$$X_{2n} \sim \mathcal{N}(z_n, \theta)$$

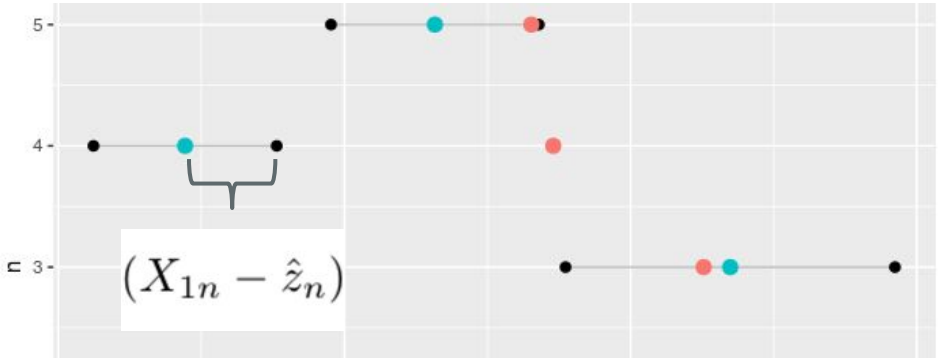
Irrespective of θ,

$$\hat{z}_n = \text{argmax}_{z_n} P(X_{1n}, X_{2n} | z_n, \theta)$$
$$= \frac{X_{1n} + X_{2n}}{2}$$



The Neyman-Scott "Paradox"

$$\hat{z}_n = \frac{X_{1n} + X_{2n}}{2}$$



$$\begin{aligned}
\hat{\theta} &= \text{argmax}_\theta P\left(X_{1n}, X_{2n} | \hat{z}_n, \theta\right) \\
&= \frac{1}{2}\left(\frac{1}{N}\sum_n \left(X_{1n} - \hat{z}_n\right)^2 + \frac{1}{N}\sum_n \left(X_{2n} - \hat{z}_n\right)^2\right) \\
&= \frac{1}{4N}\sum_n \left(X_{1n} - X_{2n}\right)^2
\end{aligned}$$

The Neyman-Scott "Paradox"

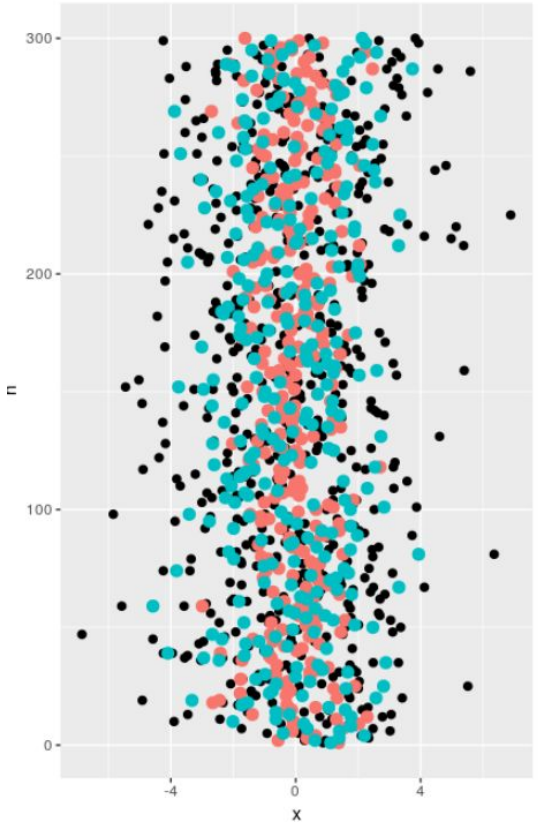
$$\hat{z}_n = \frac{X_{1n} + X_{2n}}{2} \qquad \hat{\theta} = \frac{1}{4N} \sum_n (X_{1n} - X_{2n})^2$$

What does our estimate converge to as we get more data?

$$\mathbb{E}\left[(X_{1n} - X_{2n})^2\right] = \mathbb{E}\left[\mathbb{E}\left[(X_{1n} - X_{2n})^2 \,|z_n\right]\right] = 2\theta$$

So

$$\hat{\theta} \xrightarrow[n\to\infty]{} \frac{1}{4} 2\theta = \frac{\theta}{2} \neq \theta$$

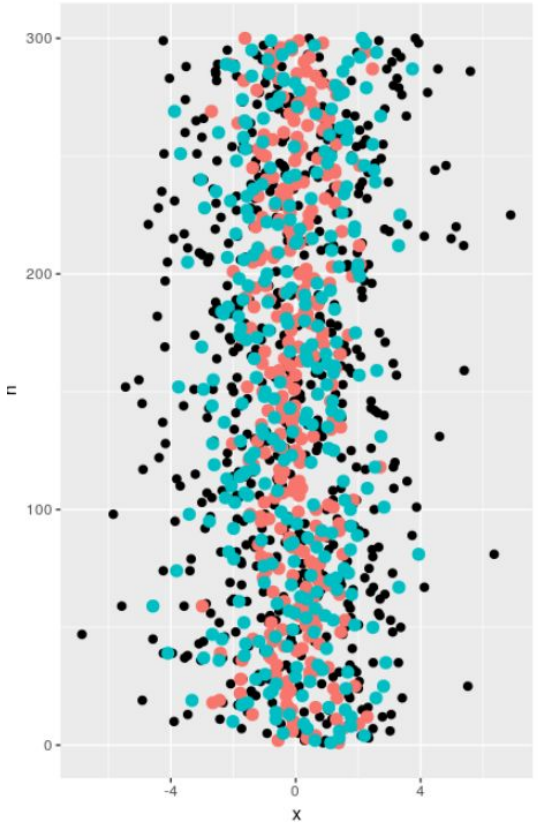


The Neyman-Scott "Paradox"

$$\hat{\theta} \xrightarrow[n \to \infty]{} \frac{1}{4} 2\theta = \frac{\theta}{2} \neq \theta$$

This is a bad estimator.

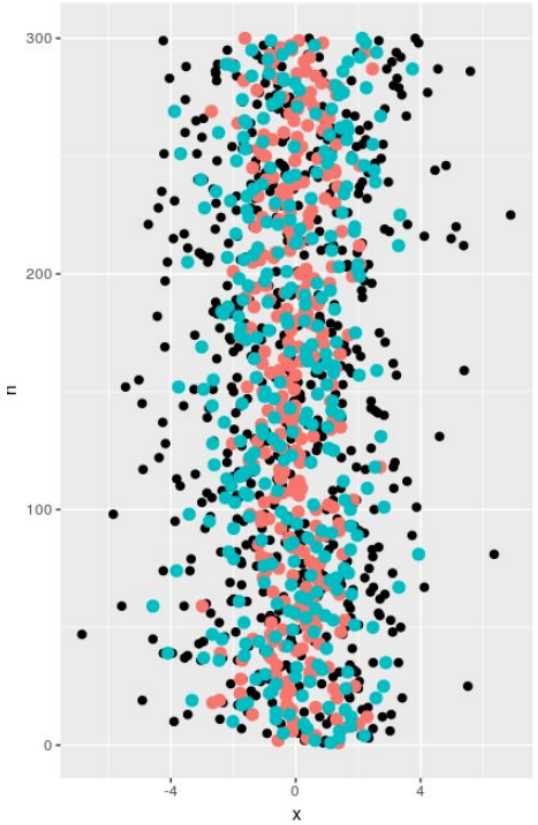
What went wrong?



The Neyman-Scott "Paradox"

Note that we estimated θ as if we knew z_n exactly, but in fact each z_n is estimated with only two data points, and is quite uncertain.

We need to take into account the uncertainty in z_n to get a good estimate of θ .



Integrate out

$$\theta = \text{Unobserved, relatively low-dimensional shared parameters}$$
$$z = (z_1, z_2, ..., z_n) = \text{Unobserved, relatively high-dimensional set of "latent" parameters or data}$$
$$X = \text{Observed data}$$

Formally, we want to "integrate out" the latent variables.

$$P(X, Z|\theta) = \text{Probability of the data and latent variables given the parameters}$$
$$P(X|\theta) = \int P(X, Z|\theta)\, dZ = \text{Marginal probability of the data given the parameters}$$

Instead of fixing *z* at some estimate, represent it as a probability distribution and calculate a solution that averages over that probability distribution.

Data *versus* parameters: doesn't matter

Are latent variables data or parameters?

$P(X, Z|\theta)$ means the latent variables are unobserved data

$P(X|Z, \theta)$ means the latent variables are extra parameters

If the Z are parameters, then in order to integrate then Z, we need to posit a probability distribution, $P(Z|\theta)$ (possibly involving extra parameters in θ). Then

$$P(X|Z, \theta) = P(X, Z|\theta)\, P(Z|\theta)$$

In the Neyman-Scott paradox, is z a parameter or data?

A resolution to the Neyman-Scott "Paradox"

Suppose we assume

$$z_n \sim \mathcal{N}(0, \theta_z)$$

Then standard facts about the bivariate normal give

$$P\left(\begin{pmatrix} X_{1n} \\ X_{2n} \end{pmatrix} | \theta, \theta_z\right) = \int P\left(\begin{pmatrix} X_{1n} \\ X_{2n} \end{pmatrix}, z_n | \theta, \theta_z\right) dz_n = \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \theta_z + \theta & \theta_z \\ \theta_z & \theta_z + \theta \end{pmatrix}\right)$$

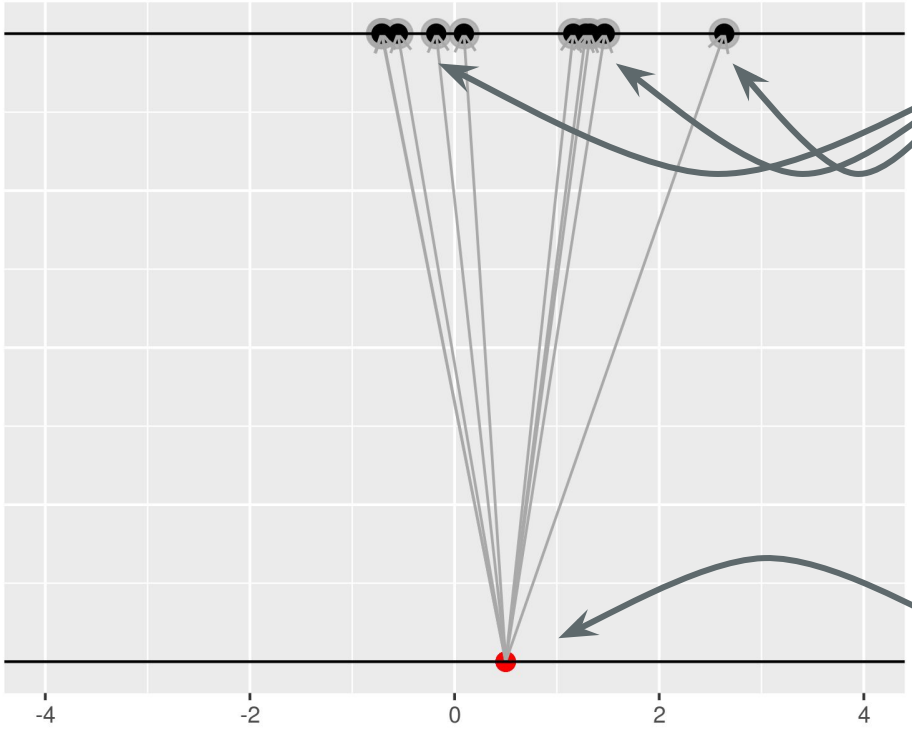
The sample covariance of a bivariate normal is consistent, so

$$X_n = \begin{pmatrix} X_{1n} \\ X_{2n} \end{pmatrix} \quad \bar{X} = \frac{1}{N}\sum_{n=1}^{N} X_n \quad \hat{\Sigma} = \frac{1}{N}\sum_{n=1}^{N} (X_n - \bar{X})(X_n - \bar{X})^T$$

$$\hat{\Sigma} \xrightarrow[n\to\infty]{} \begin{pmatrix} \theta_z + \theta & \theta_z \\ \theta_z & \theta_z + \theta \end{pmatrix} \Rightarrow \hat{\Sigma}_{11} - \hat{\Sigma}_{12} \xrightarrow[n\to\infty]{} \theta$$

Bayesian and Frequentist statistics

Parameters and data



X (data)

Probabilistic model

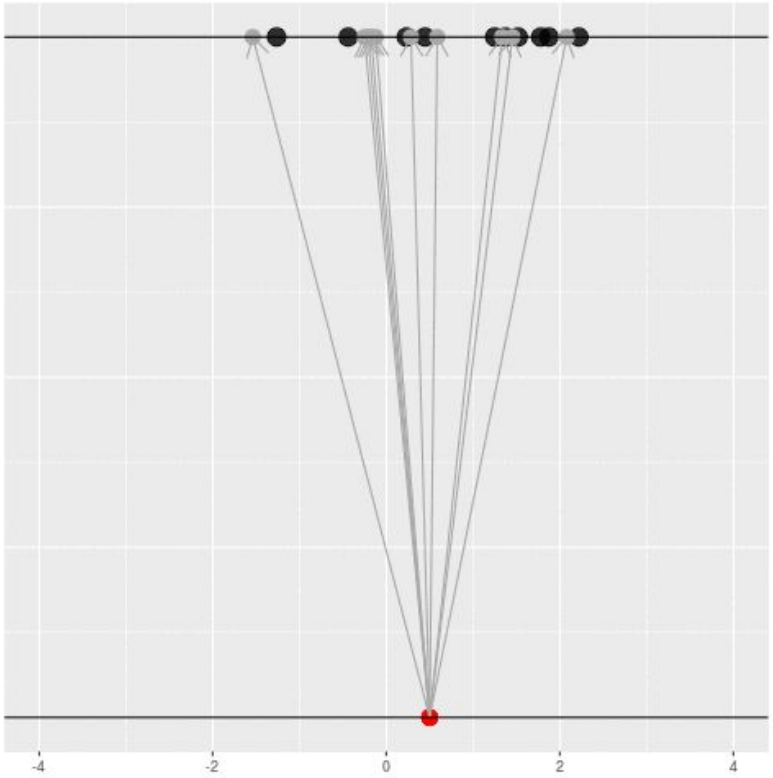
θ (parameter)

Observed

Inference

Unobserved

Frequentist approach



X (data)

Probabilistic model

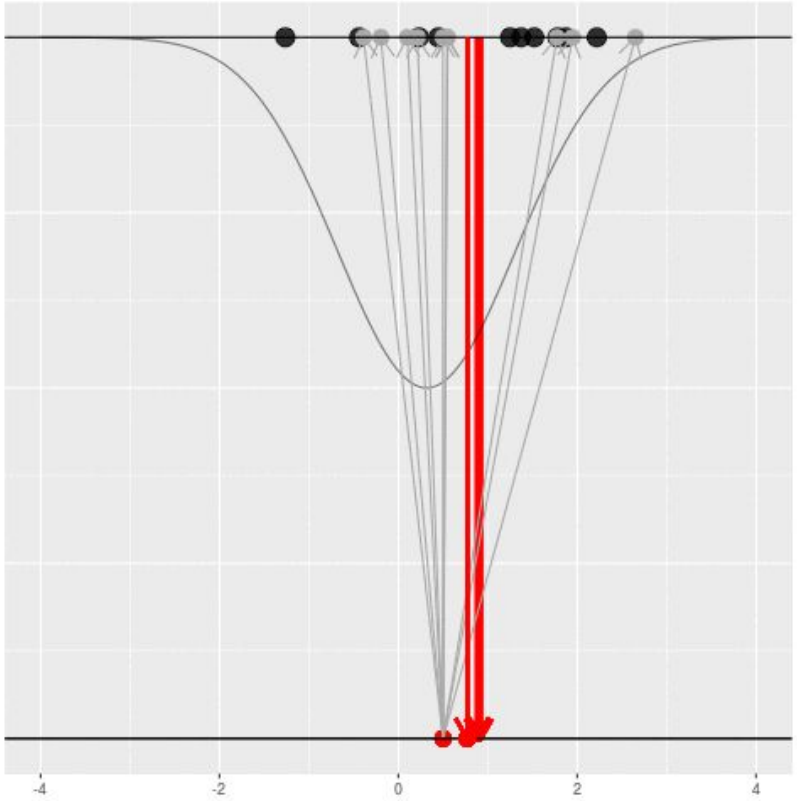
θ (parameter)

Frequentist idea:

We got the parameter indicated by the red dot and saw the dataset in black.

But the same parameter could have given us lots of other datasets.

Frequentist approach



X (data)

Probabilistic model

θ (parameter)

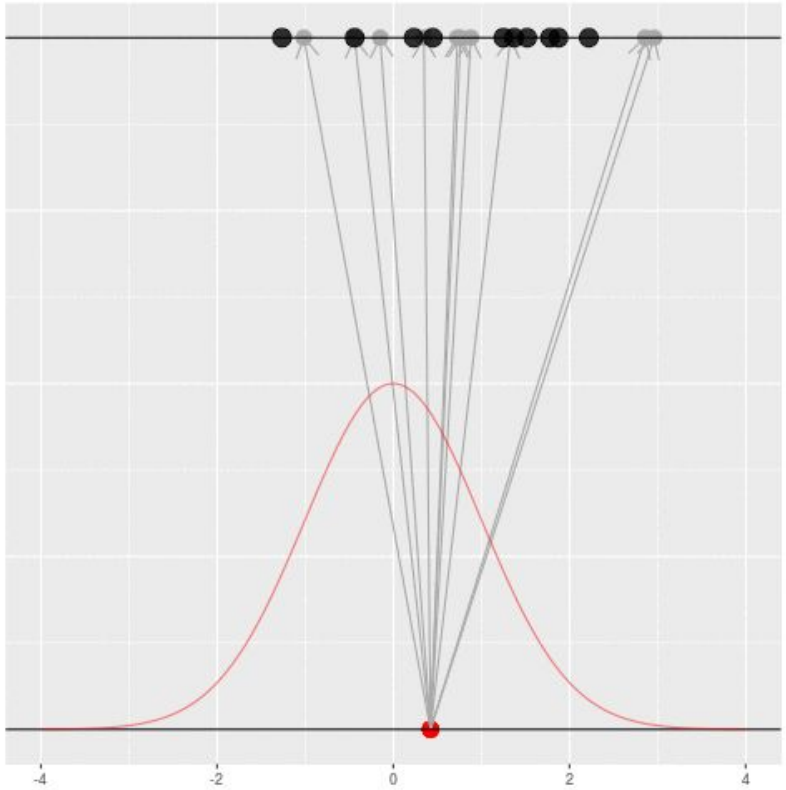
Frequentist idea:

For each dataset, we might pick some summary function and call it an “estimate”.
It will be different each time because the data will be different each time.

A typical estimate -- but not the only one -- is the value that maximizes the likelihood of the data.

We hope the estimate is usually near the true parameter in some sense.

Bayesian approach



X (data)

Probabilistic model

θ (parameter)

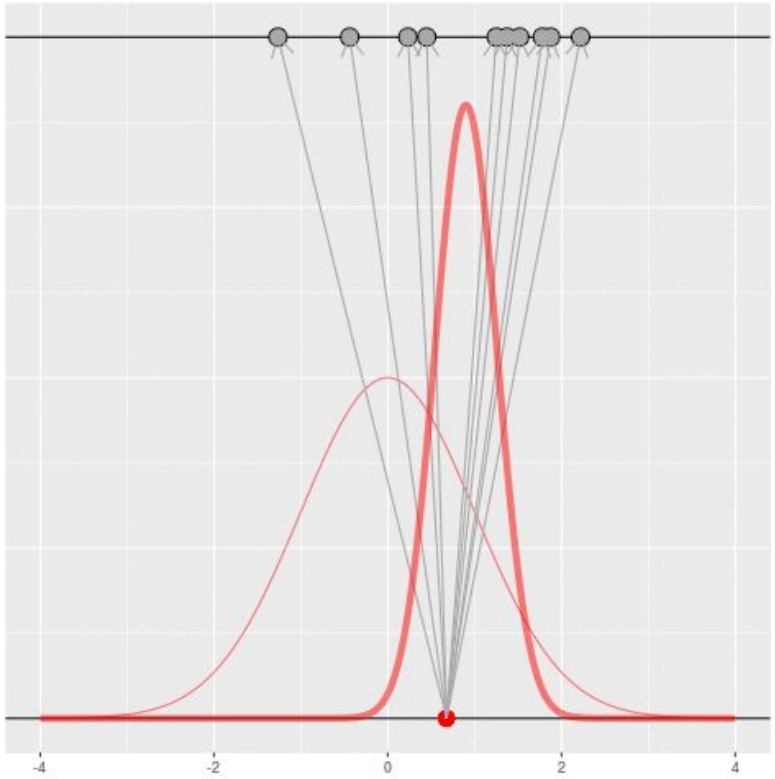
Bayesian idea:

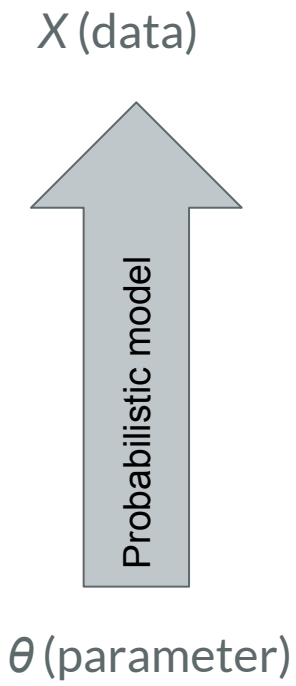
Let's imagine that the universe generates datasets by

- First drawing a parameter from some prior distribution
- and then drawing a dataset from that parameter.

Sometimes we would get the dataset we saw. Mostly we wouldn't.

Bayesian approach





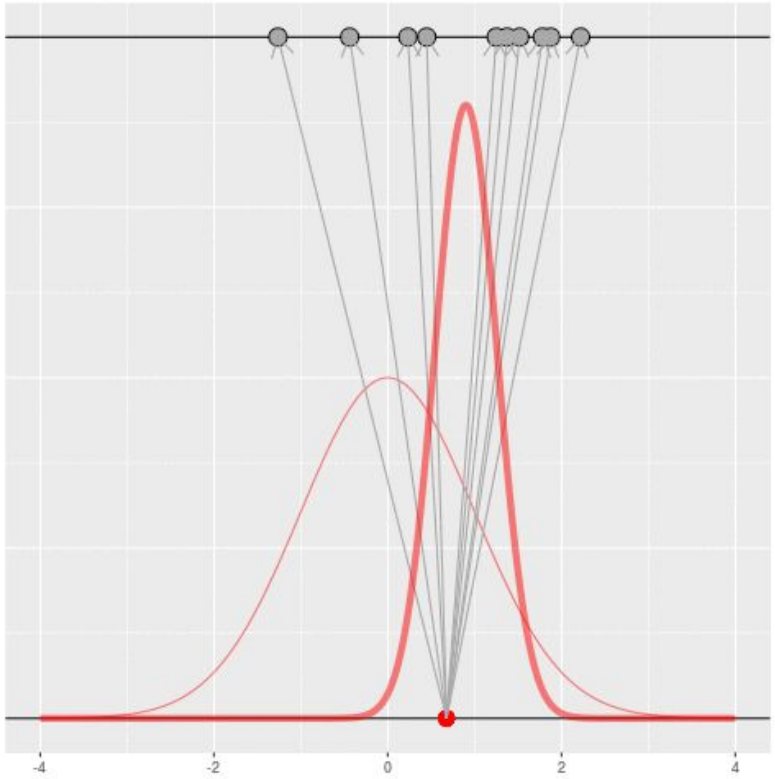
Bayesian idea:

Suppose we draw a bunch of parameters and datasets, and then throw out every pair where the data doesn't match what we observed.

The distribution of the parameters that are left represents which parameters could have given us the dataset we saw.

We hope the prior is reasonable and the model is accurate.

Bayesian approach



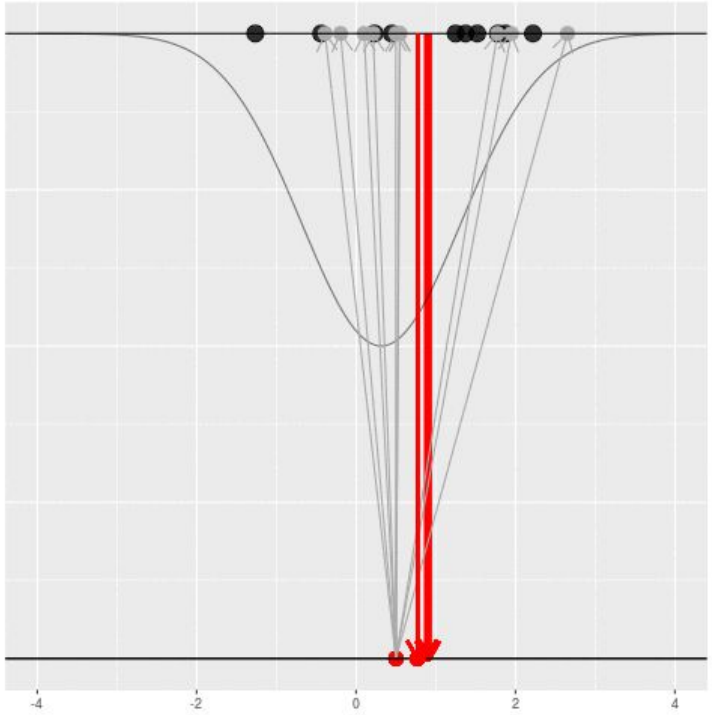
Of course, in practice you don't usually generate parameters and data hoping to get your original dataset.

Instead, you use Bayes' rule:

$$p\left(\theta|x\right) = \frac{p\left(x|\theta\right)p\left(\theta\right)}{p\left(x\right)}$$

This is intractable in general (the denominator is a problem). Turn to approximations schemes like MCMC, variational Bayes, &c.

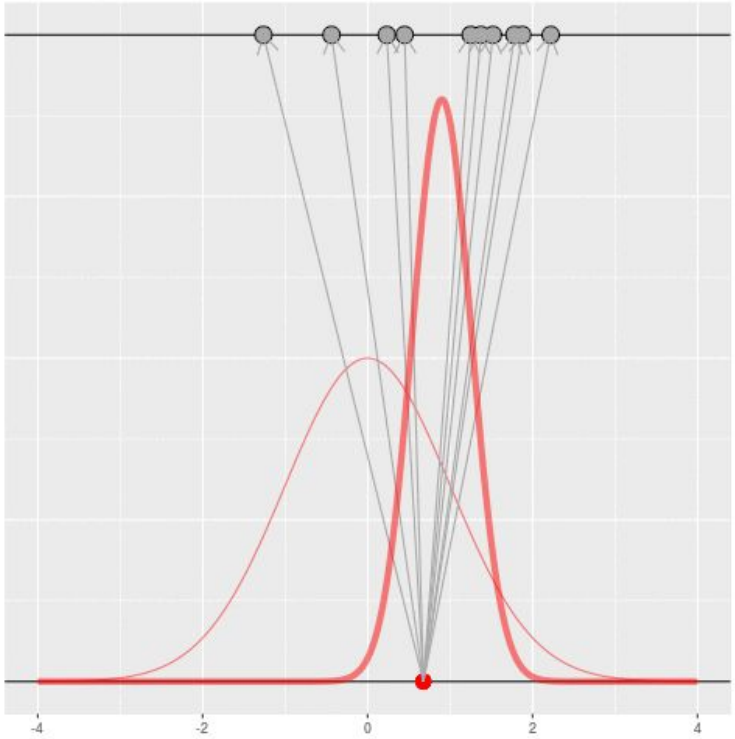
Frequentist: the parameter is fixed, data is random.
Bayes: the data is fixed, the parameter is random.



X (data)

Probabilistic model

θ (parameter)



Latent variables

Often, the frequentist estimate is

$$\hat{\theta} = \sup_{\theta} P(X|\theta) = \sup_{\theta} \int P(X, Z|\theta) = \sup_{\theta} \int P(X|Z, \theta)\, P(Z|\theta)\, dZ$$

Frequentists need to calculate this integral (the marginal likelihood), and then maximize it.

Latent variables

Often, the frequentist estimate is

$$\hat{\theta} = \sup_{\theta} P(X|\theta) = \sup_{\theta} \int P(X, Z|\theta) = \sup_{\theta} \int P(X|Z, \theta)\, P(Z|\theta)\, dZ$$

Bayesians want the posterior

$$P(\theta|X) = \frac{P(X|\theta)\, P(\theta)}{P(X)}$$

Latent variables

Often, the frequentist estimate is

$$\hat{\theta} = \sup_{\theta} P(X|\theta) = \sup_{\theta} \int P(X, Z|\theta) = \sup_{\theta} \int P(X|Z, \theta)\, P(Z|\theta)\, dZ$$

Bayesians want the posterior

$$P(\theta|X) = \frac{P(X|\theta)\, P(\theta)}{P(X)}$$

Doing posterior inference on the marginal likelihood is equivalent to marginalizing the posterior:

$$\begin{aligned} P(\theta|X) &= \frac{\left(\int P(X|Z, \theta)\, P(Z|\theta)\, dZ\right) P(\theta)}{P(X)} \\ &= \int \frac{P(X|Z, \theta)\, P(Z|\theta)\, P(\theta)}{P(X)} dZ \\ &= \int P(Z, \theta|X)\, dZ \end{aligned}$$

Latent variables

Often, the frequentist estimate is

$$\hat{\theta} = \sup_{\theta} P(X|\theta) = \sup_{\theta} \int P(X, Z|\theta) = \sup_{\theta} \int P(X|Z, \theta)\, P(Z|\theta)\, dZ$$

Bayesians want the posterior

$$P(\theta|X) = \frac{P(X|\theta)\, P(\theta)}{P(X)}$$

Doing posterior inference on the marginal likelihood is equivalent to marginalizing the posterior:

$$\begin{aligned} P(\theta|X) &= \frac{\left(\int P(X|Z, \theta)\, P(Z|\theta)\, dZ\right) P(\theta)}{P(X)} \\ &= \int \frac{P(X|Z, \theta)\, P(Z|\theta)\, P(\theta)}{P(X)} dZ \\ &= \int P(Z, \theta|X)\, dZ \end{aligned}$$

Either use the marginal likelihood in Bayes' rule

...or integrate the joint posterior (MCMC)

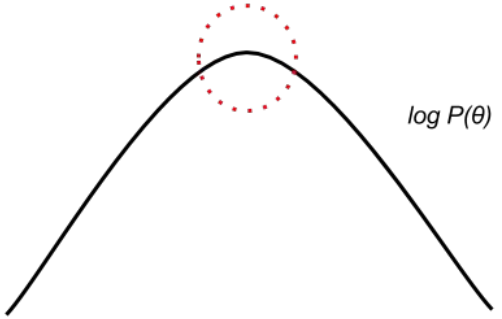
MAP and MLE: Gaussian distributions

If a distribution is Gaussian, then its log probability is quadratic, and

$$\mathbb{E}[\theta] = \hat{\theta} = \text{argmax}_\theta \log P(\theta)$$

$$\text{Cov}(\theta) = \left(\frac{\partial^2 \log P(\theta)}{\partial \theta \partial \theta^T} \bigg|_{\hat{\theta}} \right)^{-1}$$

Local information at the optimum tells you everything.



Central limit theorems

Coarsely speaking, when you have a lot of data per parameter,

- $\log P(X|\theta)$ is nearly quadratic around its maximum, $\hat{\theta}_{MLE} = \text{argmax} \log P(X|\theta)$
- $\log P(\theta|X)$ is nearly quadratic around its maximum, $\hat{\theta}_{MAP} = \text{argmax} \log P(\theta|X)$

This can be used to justify

- Estimating the mean with a maximum
- Estimating the covariance with a negative inverse Hessian

in both Bayesian and frequentist problems. For Bayesian problems, it's called the "Laplace approximation".

The MAP and the MLE

The MAP and the MLE are close when the prior is not informative relative to the data.

$$
\begin{aligned}
\hat{\theta}_{MAP} &= \text{argmax}_\theta \log P(\theta|X) \\
&= \text{argmax}_\theta \left(\log P(X|\theta) + \log P(\theta) - \log P(X)\right) \\
&\approx \text{argmax}_\theta \log P(X|\theta) \\
&= \hat{\theta}_{MLE}
\end{aligned}
$$

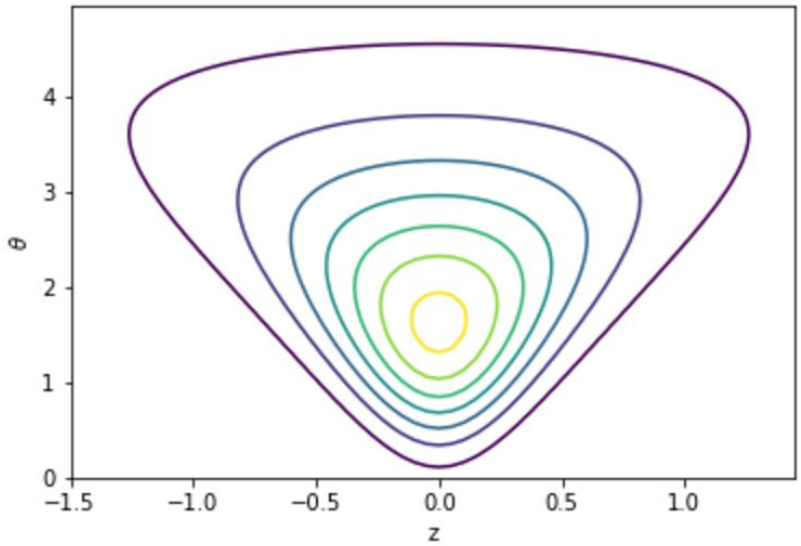
Central limit theorems

Latent variable problems may not be amenable to MAP or MLE solutions when

- There are lots of unknowns per variable
- The problem has non-linearities in the conditional distributions

In the Neyman-Scott paradox, we have both problems.

Neyman-Scott joint posterior for a single observation pair

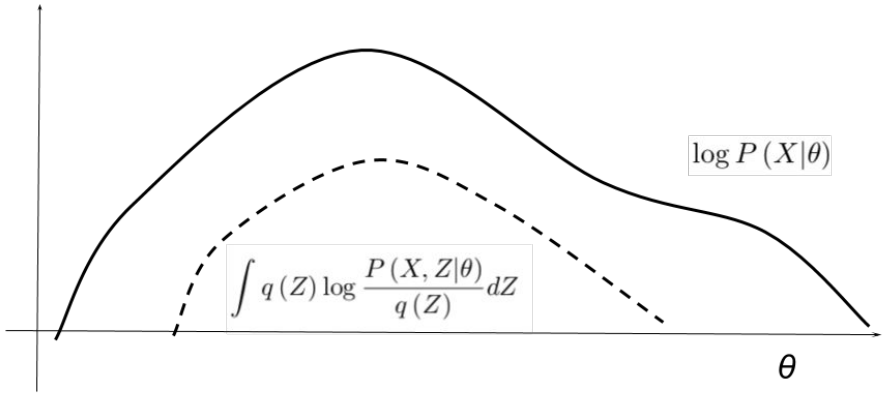


The EM Algorithm

The EM algorithm

We want to maximize $P(X|\theta)$, (or, equivalently, $\log P(X|\theta)$) but it's hard to calculate $\int P(X, Z|\theta)\, dZ$. By Jensen's inequality and concavity of the log function, for any distribution $q(Z)$,

$$\log \int P(X, Z|\theta)\, dZ = \log \int P(X, Z|\theta) \frac{q(Z)}{q(Z)} dZ \geq \int q(Z) \log \frac{P(X, Z|\theta)}{q(Z)} dZ$$

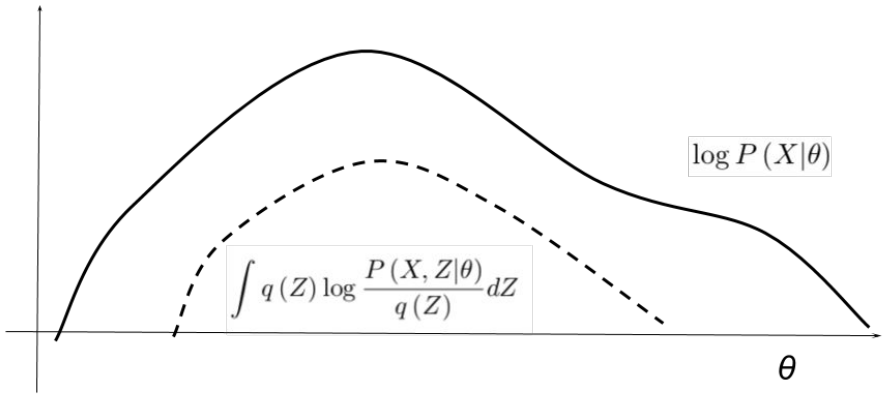


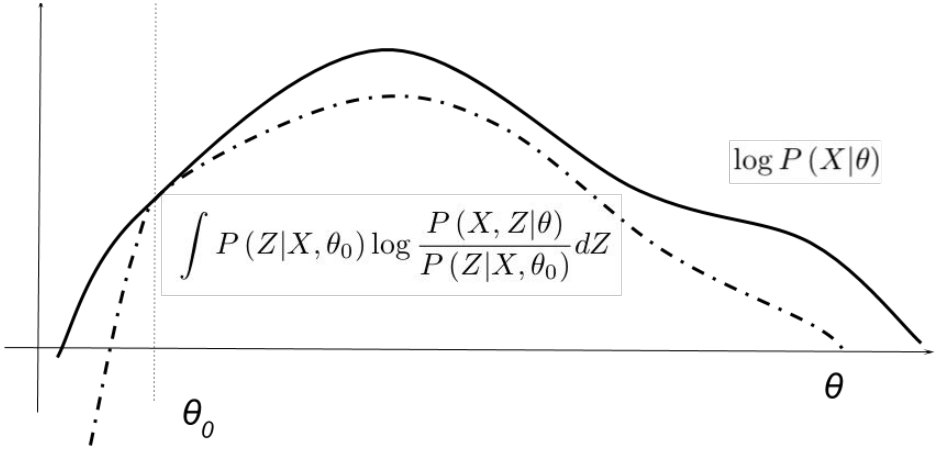
The EM algorithm

Furthermore, if for some θ_0 $q(Z) = P(Z|X,\theta_0)$ then the inequality is an equality:

$$\int P(Z|X,\theta_0)\log\frac{P(X,Z|\theta_0)}{P(Z|X,\theta_0)}dZ = \int P(Z|X,\theta_0)\log\frac{P(X,Z|\theta_0)\,P(X|\theta_0)}{P(X,Z|\theta_0)}dZ$$

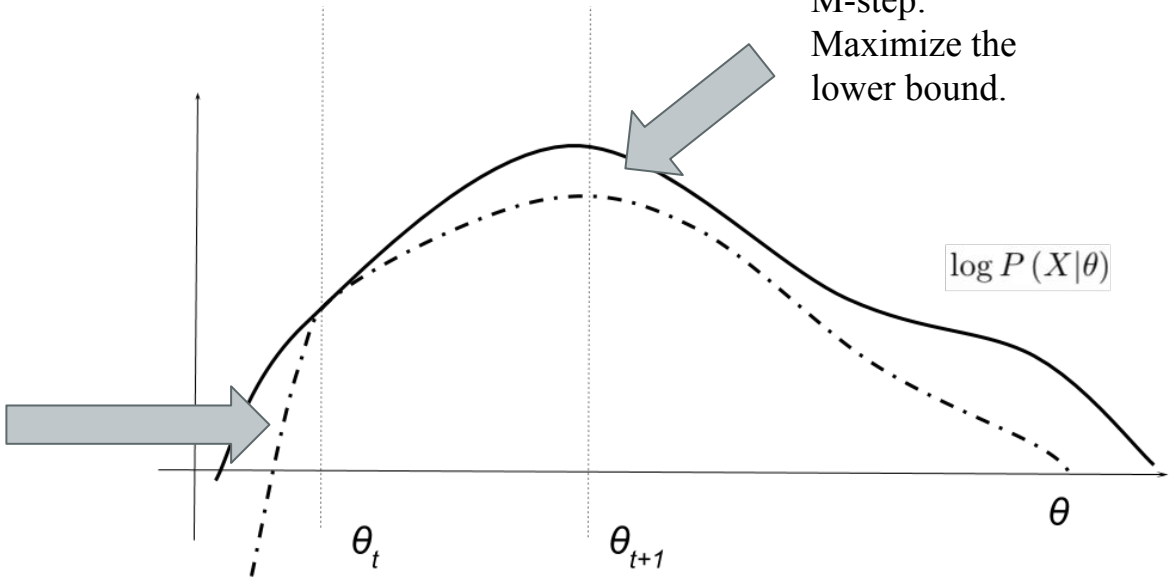
$$= \log P(X|\theta_0)\int P(Z|X,\theta_0)\,dZ = \log P(X|\theta_0)$$





The EM algorithm

1. E-step: Starting at θ_t, calculate the expectation $E(\theta) = \int P(Z|X,\theta_t) \log P(X,Z|\theta)\, dZ$
2. M-step: Optimize $\theta_{t+1} = \text{argsup} E(\theta)$
3. Repeat.

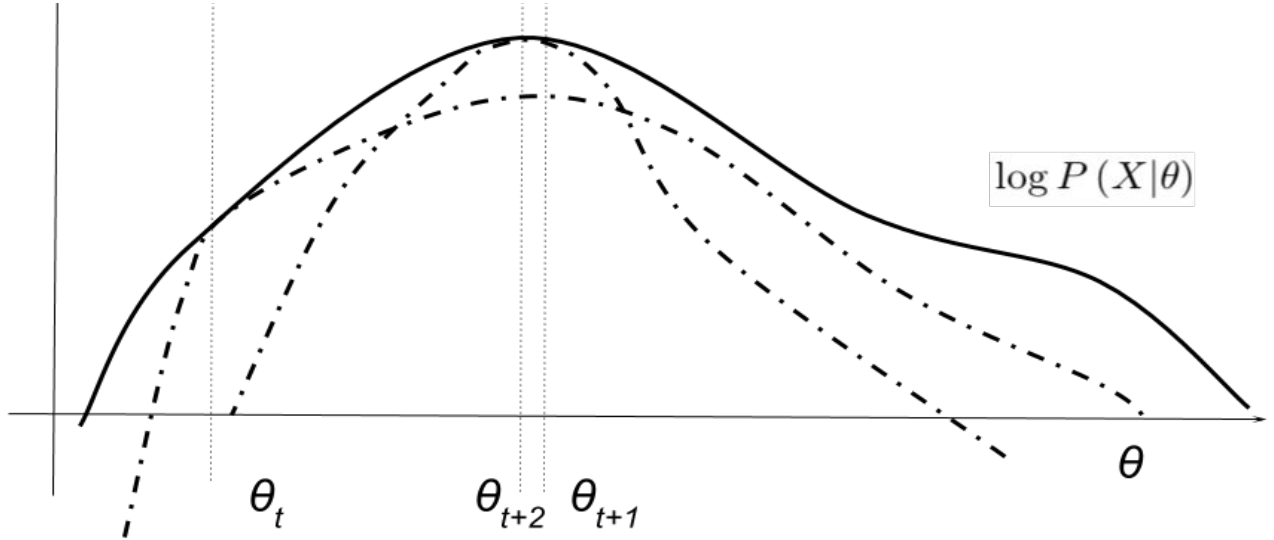


The EM algorithm

This is guaranteed to increase the marginal likelihood $\log P(\theta|X)$ since

$$\sup_\theta \int P(Z|X,\theta_t) \log P(X,Z|\theta)\, dZ = \sup_\theta \int P(Z|X,\theta_t) \log \frac{P(X,Z|\theta)}{P(Z|X,\theta_t)} dZ$$

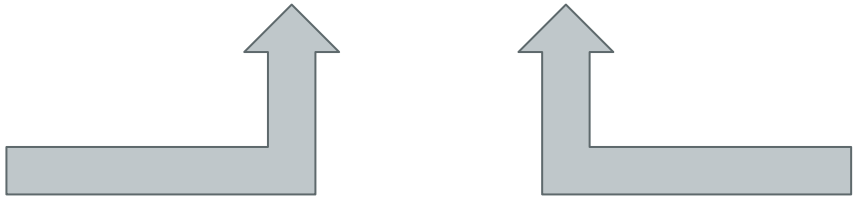
$$\geq \int P(Z|X,\theta_t) \log \frac{P(X,Z|\theta_t)}{P(Z|X,\theta_t)} dZ = \log P(\theta_t|X)$$



The EM algorithm

The EM algorithm is useful because / when:

$$\sup_\theta \int P(Z|X,\theta_t) \log P(X,Z|\theta)\, dZ$$



Conditionals are simpler than marginals

Log probabilities are simpler than probabilities (no Z in the normalizing constant!)

The Neyman-Scott paradox with EM

Recall that the probability for the Neyman-Scott observations were

$$\log P\left(X_{1n}|z_n,\theta_t\right) = -\frac{1}{2}\theta_t^{-1}\left(X_{1n}^2 - 2X_{1n}z_n + z_n^2\right) - \log\theta_t + C$$

$$\log P\left(X_{2n}|z_n,\theta_t\right) = -\frac{1}{2}\theta_t^{-1}\left(X_{2n}^2 - 2X_{2n}z_n + z_n^2\right) - \log\theta_t + C$$

$$\log P\left(X_{1n},X_{2n}|z_n,\theta_t\right) = -\frac{1}{2}\theta_t^{-1}\left(X_{1n}^2 + X_{2n}^2 - 2\left(X_{1n}+X_{2n}\right)z_n + 2z_n^2\right) + C$$

It follows that

$$P\left(z_n|X_{1n},X_{2n},\theta_t\right) = \mathcal{N}\left(\frac{X_{1n}+X_{2n}}{2},\frac{\theta_t}{2}\right)$$

The Neyman-Scott paradox with EM

It follows that

$$P\left(z_n|X_{1n},X_{2n},\theta_t\right)=\mathcal{N}\left(\frac{X_{1n}+X_{2n}}{2},\frac{\theta_t}{2}\right)$$

$$\int P\left(z_n|X_{1n},X_{2n},\theta_t\right)\log P\left(X_{1n},X_{2n},z_n|\theta\right)dZ$$

$$=-\frac{1}{2}\theta^{-1}\left(X_{1n}^2+X_{2n}^2-2\left(X_{1n}+X_{2n}\right)\mathbb{E}_q\left[z_n\right]+2\mathbb{E}_q\left[z_n^2\right]\right)-\frac{1}{2}\log\theta$$

$$=-\frac{1}{2}\theta^{-1}\left(X_{1n}^2+X_{2n}^2-\left(X_{1n}+X_{2n}\right)^2+\frac{1}{2}\left(X_{1n}+X_{2n}\right)^2+\theta_t\right)-\frac{1}{2}\log\theta$$

The Neyman-Scott paradox with EM

$$
\begin{aligned}
&\int P\left(z_n|X_{1n},X_{2n},\theta_t\right)\log P\left(X_{1n},X_{2n},z_n|\theta\right)dZ \\
&= -\frac{1}{2}\theta^{-1}\left(X_{1n}^2+X_{2n}^2-2\left(X_{1n}+X_{2n}\right)\mathbb{E}_q\left[z_n\right]+2\mathbb{E}_q\left[z_n^2\right]\right)-\frac{1}{2}\log\theta \\
&= -\frac{1}{2}\theta^{-1}\left(X_{1n}^2+X_{2n}^2-\left(X_{1n}+X_{2n}\right)^2+\frac{1}{2}\left(X_{1n}+X_{2n}\right)^2+\theta_t\right)-\frac{1}{2}\log\theta \\
&= -\frac{1}{2}\theta^{-1}\left(X_{1n}^2+X_{2n}^2-\frac{1}{2}\left(X_{1n}+X_{2n}\right)^2+\theta_t\right)-\frac{1}{2}\log\theta \\
&= -\frac{1}{2}\theta^{-1}\left(X_{1n}^2+X_{2n}^2-\frac{1}{2}\left(X_{1n}^2+X_{2n}^2+2X_{1n}X_{2n}\right)+\theta_t\right)-\frac{1}{2}\log\theta \\
&= -\frac{1}{4}\theta^{-1}\left(X_{1n}^2+X_{2n}^2-2X_{1n}X_{2n}+\theta_t\right)-\frac{1}{2}\log\theta \\
&= -\frac{1}{4}\theta^{-1}\left(\left(X_{1n}-X_{2n}\right)^2+\theta_t\right)-\frac{1}{2}\log\theta
\end{aligned}
$$

The Neyman-Scott paradox with EM

The next step is given by

$$\theta_{t+1} = \text{argsup}_\theta - \frac{1}{2}\theta^{-1}\sum_{n=1}^{N}\left((X_{1n} - X_{2n})^2 + \theta_t\right) - N\frac{1}{2}\log\theta$$

$$0 = \frac{1}{2}\theta_{t+1}^{-2}\sum_{n=1}^{N}\left((X_{1n} - X_{2n})^2 + \theta_t\right) - N\frac{1}{2}\theta_{t+1}^{-1}$$

$$\theta_{t+1} = \frac{1}{N}\sum_{n=1}^{N}\left((X_{1n} - X_{2n})^2 + \theta_t\right)$$

The Neyman-Scott paradox with EM

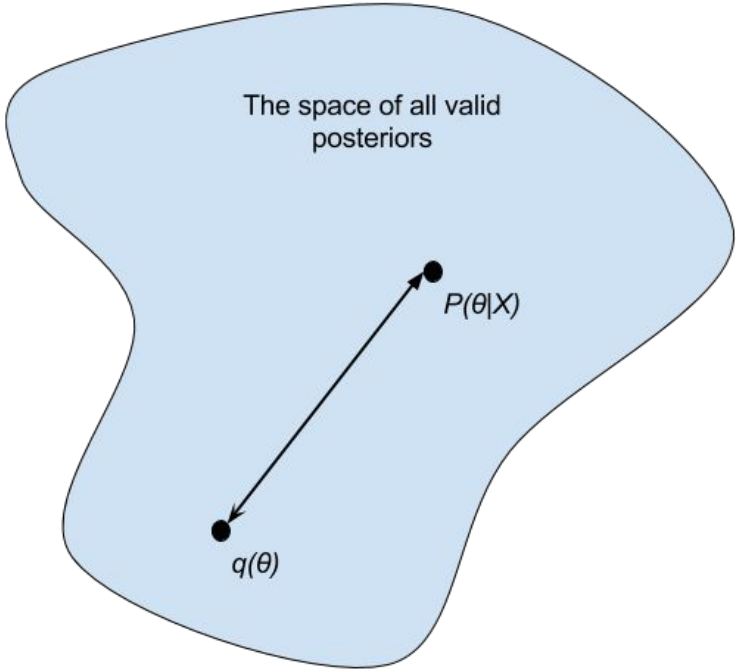
A fixed point is

$$\hat{\theta} = \frac{1}{N} \sum_{n=1}^{N} \left((X_{1n} - X_{2n})^2 + \hat{\theta} \right)$$

$$\hat{\theta} = \frac{1}{2N} \sum_{n=1}^{N} (X_{1n} - X_{2n})^2 \xrightarrow[n \to \infty]{} \theta$$

Variational Bayes

Variational Bayes writ large



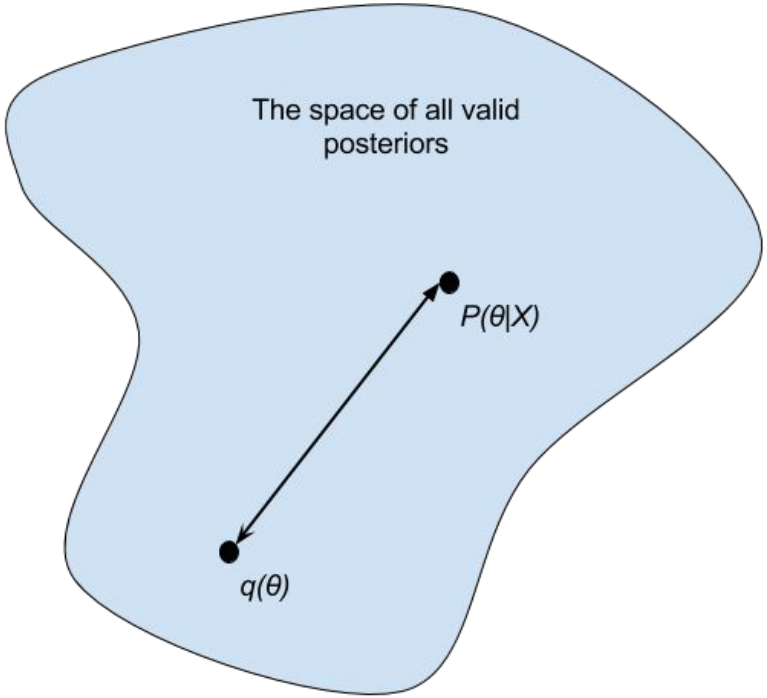
A tautology:

Suppose we have some notion of the "divergence" between two distributions.

The posterior is the the distribution closest to the posterior amongst all potentially valid distributions.

Variational Inference: A Review for Statisticians
David M Blei, Alp Kucukelbir, Jon D McAuliffe (2016)

KL divergence



We will use “Kullback-Liebler” (KL) divergence.

$$KL\left(q\left(\theta\right)||P\left(\theta|X\right)\right) = \int q\left(\theta\right) \log \frac{P\left(\theta|X\right)}{q\left(\theta\right)} d\theta$$

“Divergence”, not “distance”, because:

$$KL\left(q\left(\theta\right)||P\left(\theta|X\right)\right) \geq 0$$

$$KL\left(q\left(\theta\right)||P\left(\theta|X\right)\right) = 0 \Leftrightarrow q\left(\theta\right) = P\left(\theta|X\right)$$

but

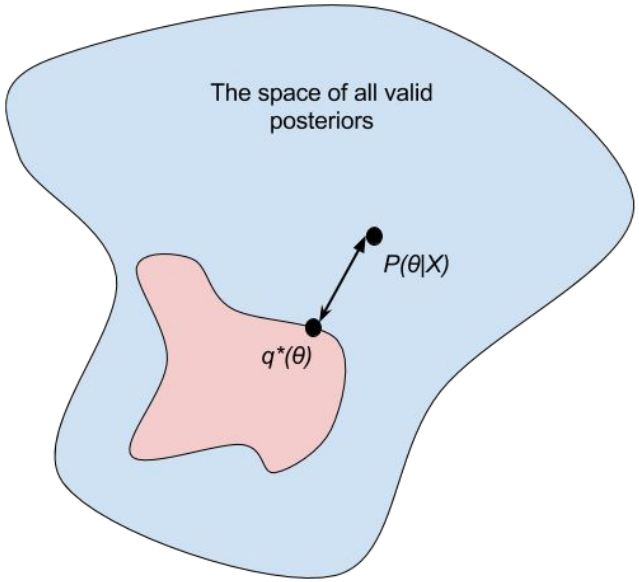
$$KL\left(q\left(\theta\right)||P\left(\theta|X\right)\right) \neq KL\left(P\left(\theta|X\right)||q\left(\theta\right)\right)$$

Variational Inference: A Review for Statisticians
David M Blei, Alp Kucukelbir, Jon D McAuliffe (2016)

Why is KL divergence practically nice?

$$
\begin{aligned}
P(\theta|X) &= \text{argmin}_q KL(q(\theta)\,||P(\theta|X)) \\
&= \text{argmin}_q \int q(\theta) \log \frac{q(\theta)}{P(\theta|X)} d\theta \\
&= \text{argmin}_q \left\{ \int q(\theta) \log q(\theta)\, d\theta - \int q(\theta) \log P(\theta, X)\, P(\theta)\, d\theta - P(X) \right\} \\
&= \text{argmax}_q \left\{ \underbrace{-\int q(\theta) \log q(\theta)\, d\theta}_{\text{Entropy of approximation}} + \underbrace{\int q(\theta) \log P(\theta, X)\, P(\theta)\, d\theta}_{\text{Data fit (without the normalizing constant!)}} \right\}
\end{aligned}
$$

Variational Bayes

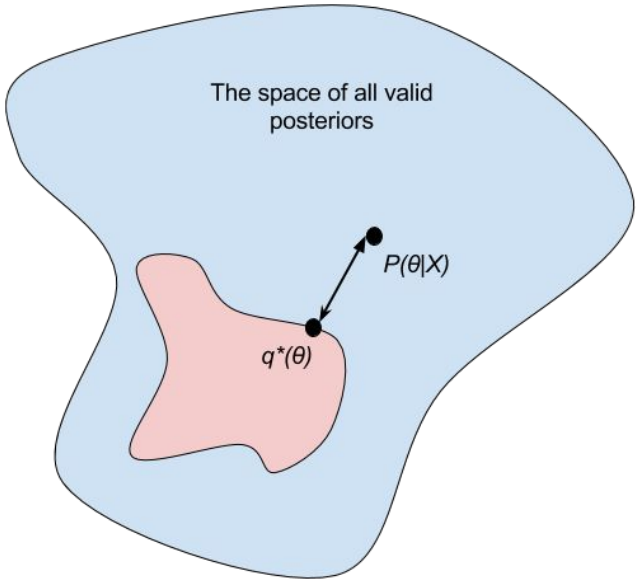


General entropies are hard to calculate. So limit the approximation to tractable distributions:

$$\mathcal{Q} = \{q \text{ that are tractable in some way}\}$$
$$q^*(\theta) = \text{argmin}_{q \in \mathcal{Q}} KL(q(\theta) || P(\theta|X))$$

Variational Inference: A Review for Statisticians
David M Blei, Alp Kucukelbir, Jon D McAuliffe (2016)

Mean field variational Bayes



The "mean field" (MFVB) approximation uses factorizing distributions:

$$\mathcal{Q} = \left\{ q(\theta) = \prod_k q(\theta_k) \right\}$$

$$q^*(\theta) = \text{argmin}_{q \in \mathcal{Q}} KL(q(\theta) || P(\theta|X))$$

Note: this is not an "assumption", because it is ridiculous.

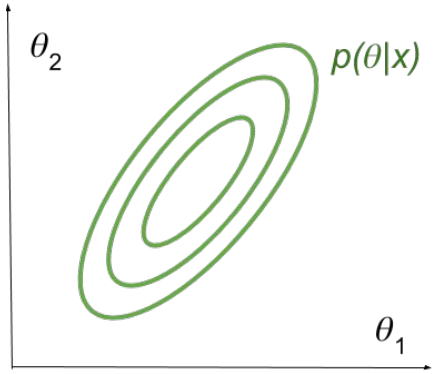
Variational Inference: A Review for Statisticians
David M Blei, Alp Kucukelbir, Jon D McAuliffe (2016)

Variational Bayes - a simple example.

A bivariate normal posterior.

A mean field normal approximation.

$$\mathcal{Q} = \left\{ q(\theta) = \mathcal{N}\left(\theta_1; \mu_1, \sigma_1^2\right) \mathcal{N}\left(\theta_2; \mu_2, \sigma_2^2\right) \right\}$$
$$\eta_1 = \left(\mu_1, \sigma_1^2\right)$$
$$\eta_2 = \left(\mu_2, \sigma_2^2\right)$$
$$\eta = \left(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2\right)$$



Variational Inference: A Review for Statisticians
David M Blei, Alp Kucukelbir, Jon D McAuliffe (2016)

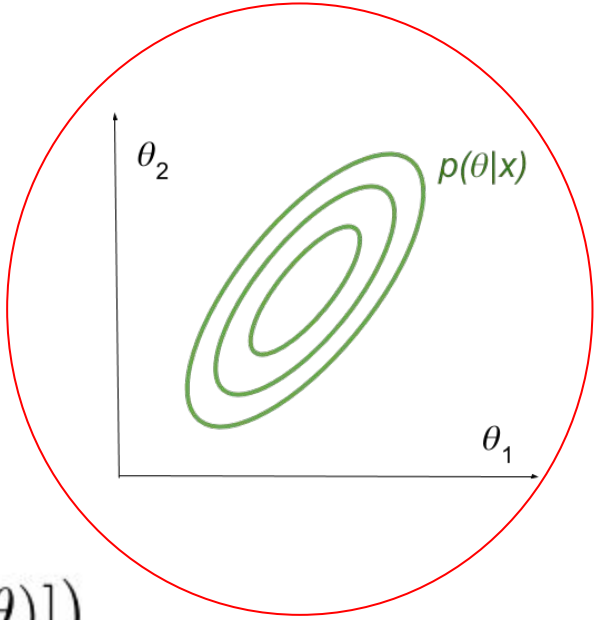
Variational Bayes - a simple example.

$$\mathcal{Q} = \left\{ q(\theta) = \mathcal{N}\left(\theta_1; \mu_1, \sigma_1^2\right) \mathcal{N}\left(\theta_2; \mu_2, \sigma_2^2\right) \right\}$$

$$\eta_1 = \left(\mu_1, \sigma_1^2\right)$$

$$\eta_2 = \left(\mu_2, \sigma_2^2\right)$$

$$\eta = \left(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2\right)$$



$$\operatorname{argmin}_{q \in \mathcal{Q}} \left(\mathbb{E}_{q(\theta)} \left[\log q(\theta) \right] - \mathbb{E}_{q(\theta)} \left[\log \quad (\theta) \right] \right)$$

Variational Inference: A Review for Statisticians
David M Blei, Alp Kucukelbir, Jon D McAuliffe (2016)

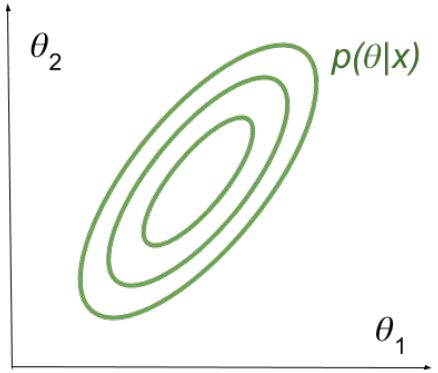
Variational Bayes - a simple example.

$$\mathcal{Q} = \left\{ q(\theta) = \mathcal{N}\left(\theta_1; \mu_1, \sigma_1^2\right) \mathcal{N}\left(\theta_2; \mu_2, \sigma_2^2\right) \right\}$$

$$\eta_1 = \left(\mu_1, \sigma_1^2\right)$$

$$\eta_2 = \left(\mu_2, \sigma_2^2\right)$$

$$\eta = \left(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2\right)$$



$$\operatorname{argmin}_{q \in \mathcal{Q}} \left(\mathbb{E}_{q(\theta)} [\log \theta)] - \mathbb{E}_{q(\theta)} \left[\log p_\alpha^x(\theta) \right] \right)$$

Variational Inference: A Review for Statisticians
David M Blei, Alp Kucukelbir, Jon D McAuliffe (2016)

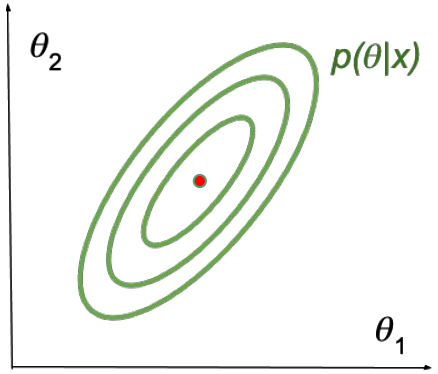
Variational Bayes - a simple example.

$$\mathcal{Q} = \left\{ q(\theta) = \mathcal{N}\left(\theta_1; \mu_1, \sigma_1^2\right) \mathcal{N}\left(\theta_2; \mu_2, \sigma_2^2\right) \right\}$$

$$\eta_1 = \left(\mu_1, \sigma_1^2\right)$$

$$\eta_2 = \left(\mu_2, \sigma_2^2\right)$$

$$\eta = \left(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2\right)$$

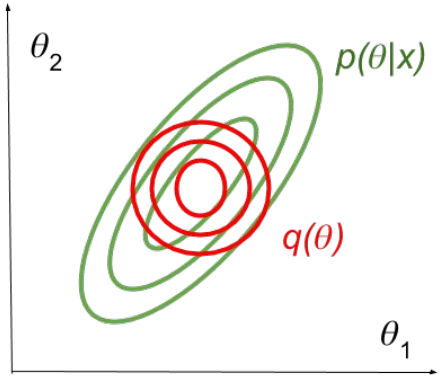


$$\operatorname{argmin}_{q \in \mathcal{Q}} \left(\mathbb{E}_{q(\theta)} [\log \theta)] - \mathbb{E}_{q(\theta)} \left[\log p_\alpha^x(\theta) \right] \right)$$

Variational Inference: A Review for Statisticians
David M Blei, Alp Kucukelbir, Jon D McAuliffe (2016)

Variational Bayes - a simple example.

$$\mathcal{Q} = \left\{ q(\theta) = \mathcal{N}\left(\theta_1; \mu_1, \sigma_1^2\right) \mathcal{N}\left(\theta_2; \mu_2, \sigma_2^2\right) \right\}$$
$$\eta_1 = \left(\mu_1, \sigma_1^2\right)$$
$$\eta_2 = \left(\mu_2, \sigma_2^2\right)$$
$$\eta = \left(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2\right)$$



$$\operatorname{argmin}_{q \in \mathcal{Q}} \left(\mathbb{E}_{q(\theta)} \left[\log q(\theta) \right] - \mathbb{E}_{q(\theta)} \left[\log p_\alpha^x(\theta) \right] \right)$$

Variational Inference: A Review for Statisticians
David M Blei, Alp Kucukelbir, Jon D McAuliffe (2016)

Variational Bayes and the EM algorithm

Suppose we assume that

$$\mathcal{Q} = \{q(\theta, Z) = \delta(\theta - \theta_0)\, q(Z)\}$$
$$q(\theta, Z) = \text{argmin}_q KL(q||P(\theta, Z|X))$$

Then we recover the EM algorithm as a special case.

The Variational Bayesian EM Algorithm for Incomplete Data:
Matthew Beal and Zoubin Ghahramani (2003)

Variational Bayes and the EM algorithm

Updating $q(Z)$, keeping $q(\theta)$ fixed, is the E-step:

$$
\begin{aligned}
q^*(Z) &= \text{argmin}_{q(Z)} KL\left(q(Z)\,\delta(\theta-\theta_0)\,||P(\theta,Z|X)\right) \\
&= \text{argmin}_{q(Z)} \int q(Z)\,\delta(\theta-\theta_0)\,\frac{q(Z)\,\delta(\theta-\theta_0)}{\log P(\theta,Z|X)}\,dZ\,d\theta \\
&= \text{argmin}_{q(Z)} \int q(Z)\,\frac{q(Z)}{\log P(\theta_0,Z|X)}\,dZ \\
&= \text{argmax}_{q(Z)} \int q(Z)\,\frac{\log P(\theta_0,Z|X)}{q(Z)}\,dZ \\
&= \text{argmax}_{q(Z)} \int q(Z)\,\frac{\log P(Z,X|\theta_0)}{q(Z)}\,dZ \\
&= P(Z|\theta_0,X)
\end{aligned}
$$

The Variational Bayesian EM Algorithm for Incomplete Data:
Matthew Beal and Zoubin Ghahramani (2003)

Variational Bayes and the EM algorithm

Updating $q(\theta)$, keeping $q(Z)$ fixed, is the M-step:

$$\begin{aligned}\theta_0 &= \text{argmin}_{\theta_0} KL\left(q\left(Z\right)\delta\left(\theta-\theta_0\right)||P\left(\theta,Z|X\right)\right)\\ &= \text{argmax}_{\theta_0}\int q\left(Z\right)\log P\left(\theta_0,Z|X\right)dZ\\ &= \text{argmax}_{\theta_0}\int q\left(Z\right)\log P\left(X,Z|\theta_0\right)P\left(\theta_0\right)dZ\end{aligned}$$

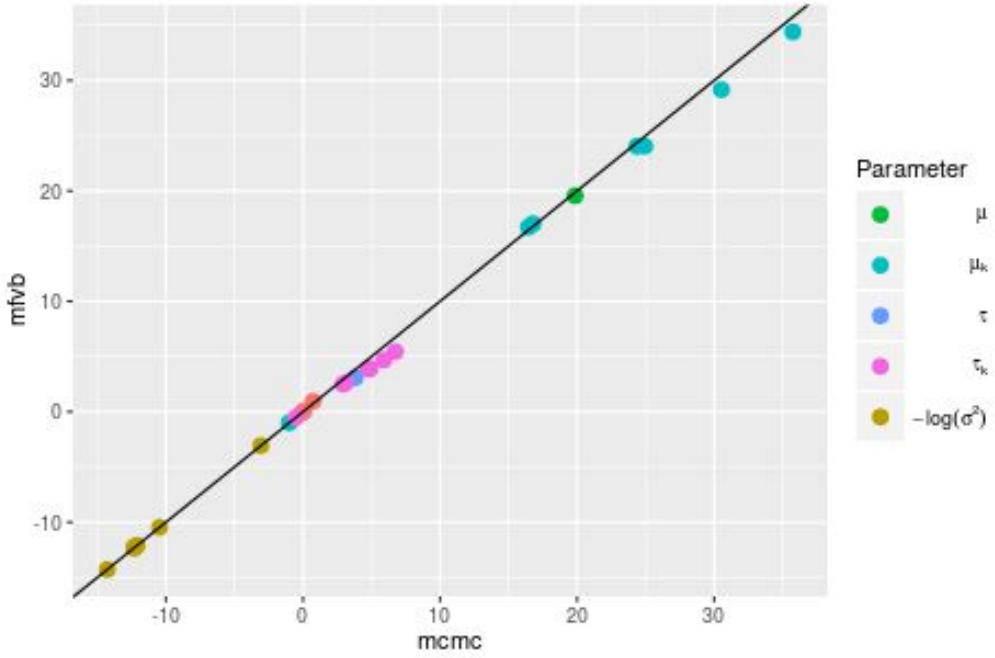
The Variational Bayesian EM Algorithm for Incomplete Data:
Matthew Beal and Zoubin Ghahramani (2003)

Why use variational Bayes?

It's very fast. On a hierarchical model from our paper:

```
MCMC time (with Stan): 45 minutes
VB time:               52 seconds
```

Often, the variational posterior means match MCMC quite closely.



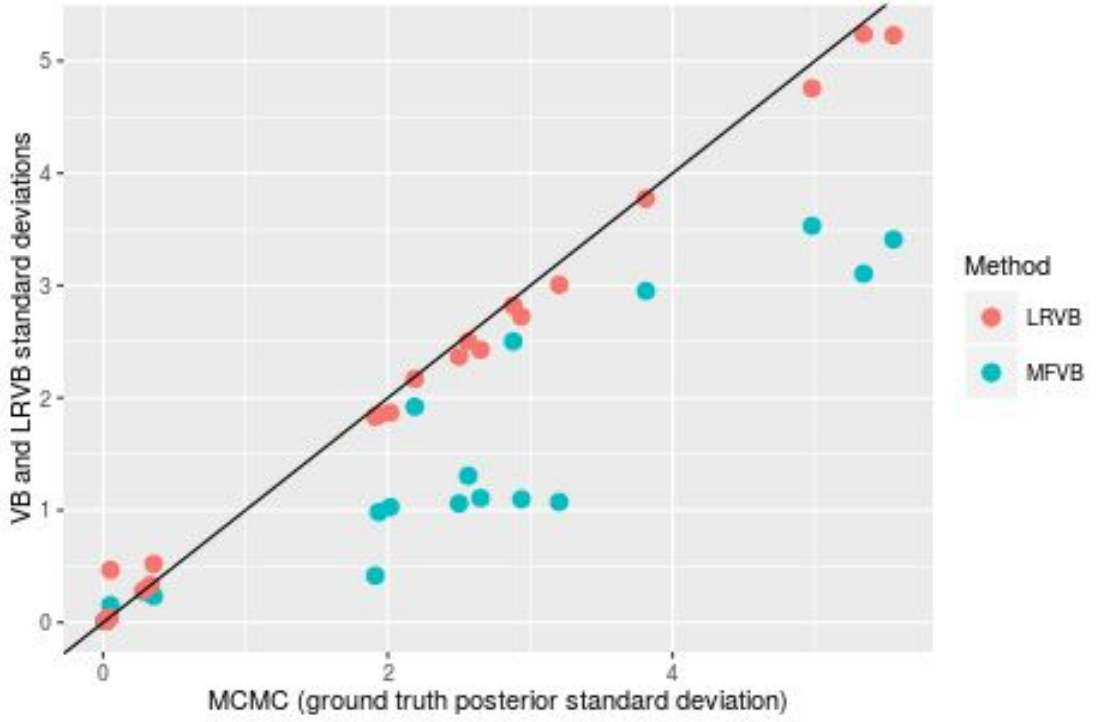
Linear Response Methods for Accurate Covariance Estimates from Mean Field Variational Bayes: Ryan Giordano, Tamara Broderick, Michael Jordan (2015)

Why not use variational Bayes?

If MFVB is wrong, it's hard to know how wrong it is without running MCMC anyway.

Historically, MFVB was not used for inference because it tends to underestimate posterior variance.

(However, we have a fix.)



Linear Response Methods for Accurate Covariance Estimates from Mean Field Variational Bayes: Ryan Giordano, Tamara Broderick, Michael Jordan (2015)

MFVB and Neyman-Scott

$$X_i, Y_i \sim \mathcal{N}\left(\alpha_i, \nu^{-1}\right)$$

$$\begin{aligned}\log P\left(X_i, Y_i | \alpha_i, \nu\right) &= -\frac{\nu}{2}\left(X_i - \alpha_i\right)^2 - \frac{\nu}{2}\left(Y_i - \alpha_i\right)^2 + \frac{2}{2}\log\nu \\ &= -\frac{\nu}{2}\left(X_i^2 + Y_i^2 - 2\left(X_i + Y_i\right)\alpha_i + 2\alpha_i^2\right) + \log\nu\end{aligned}$$

$$\log P\left(X, Y | \alpha, \nu\right) = -\frac{\nu}{2}\left(\sum_{i=1}^{N}\left(X_i^2 + Y_i^2\right) - 2\sum_{i=1}^{N}\left(X_i + Y_i\right)\alpha_i + 2\sum_{i=1}^{N}\alpha_i^2\right) + N\log\nu$$

$$q^t\left(\alpha_i\right) \leftarrow \mathcal{N}\left(\frac{X_i + Y_i}{2}, \frac{1}{2\mathbb{E}^t\left[\nu\right]}\right)$$

$$\begin{aligned}\mathbb{E}_{q^t(\alpha)}\left[\log P\left(X_i, Y_i | \alpha_i, \nu\right)\right] &= -\frac{\nu}{2}\left(\left(X_i^2 + Y_i^2\right) - 2\left(X_i + Y_i\right)\mathbb{E}\left[\alpha_i\right] + 2\mathbb{E}\left[\alpha_i^2\right]\right) + \log\nu \\ &= -\frac{\nu}{2}\left(\left(X_i^2 + Y_i^2\right) - 2\left(X_i + Y_i\right)\mathbb{E}\left[\alpha_i\right] + 2\left(\mathbb{E}\left[\alpha_i\right]^2 + \mathrm{Var}\left(\alpha_i\right)\right)\right) + \log\nu\end{aligned}$$

MFVB and Neyman-Scott

$$
\begin{aligned}
\mathbb{E}_{q^t(\alpha)}\left[\log P\left(X_i, Y_i | \alpha_i, \nu\right)\right] &= -\frac{\nu}{2}\left(\left(X_i^2 + Y_i^2\right) - \left(X_i + Y_i\right)^2 + 2\left(\frac{1}{4}\left(X_i + Y_i\right)^2 + \frac{1}{2\mathbb{E}^t\left[\nu\right]}\right)\right) + \log\nu \\
&= -\frac{\nu}{2}\left(\left(X_i^2 + Y_i^2\right) - \left(X_i + Y_i\right)^2 + \frac{1}{2}\left(X_i + Y_i\right)^2 + \frac{1}{\mathbb{E}^t\left[\nu\right]}\right) + \log\nu \\
&= -\frac{\nu}{2}\left(X_i^2 + Y_i^2 - \frac{1}{2}X_i^2 - \frac{1}{2}Y_i^2 - X_iY_i + \frac{1}{\mathbb{E}^t\left[\nu\right]}\right) + \log\nu \\
&= -\frac{\nu}{2}\left(\frac{1}{2}\left(X_i^2 + Y_i^2 - 2X_iY_i\right) + \frac{1}{\mathbb{E}^t\left[\nu\right]}\right) + \log\nu \\
&= -\frac{\nu}{2}\left(\frac{1}{2}\left(X_i - Y_i\right)^2 + \frac{1}{\mathbb{E}^t\left[\nu\right]}\right) + \log\nu
\end{aligned}
$$

MFVB and Neyman-Scott

Define $S = \sum_i (X_i - Y_i)^2$. Then the update for $q(\nu)$ i

$$q^{t+1}(\nu) \leftarrow \text{Gamma}\left(N+1, \frac{S}{4} + \frac{1}{2\mathbb{E}^t[\nu]}\right)$$

$$\mathbb{E}^{t+1}[\nu] = \frac{N+1}{\frac{S}{4} + \frac{1}{2\mathbb{E}^t[\nu]}} \Rightarrow \text{the optimum satisfies}$$

$$\mathbb{E}[\nu] = \frac{N+1}{\frac{S}{4} + \frac{1}{2\mathbb{E}[\nu]}} \Rightarrow$$

$$1 = \frac{N+1}{\frac{S}{2}\mathbb{E}[\nu] + 1} \Rightarrow$$

$$\mathbb{E}[\nu] = \frac{2N}{S}$$

MFVB and Neyman-Scott

Because

$$\begin{aligned}\mathbb{E}[S] &= \mathbb{E}\left[\sum_i (X_i - \alpha_i + \alpha_i - Y_i)^2\right] \\ &= \mathbb{E}\left[\sum_i (X_i - \alpha_i) + (Y_i - \alpha_i)^2\right] \\ &= 2N\nu_0^{-1}\end{aligned}$$

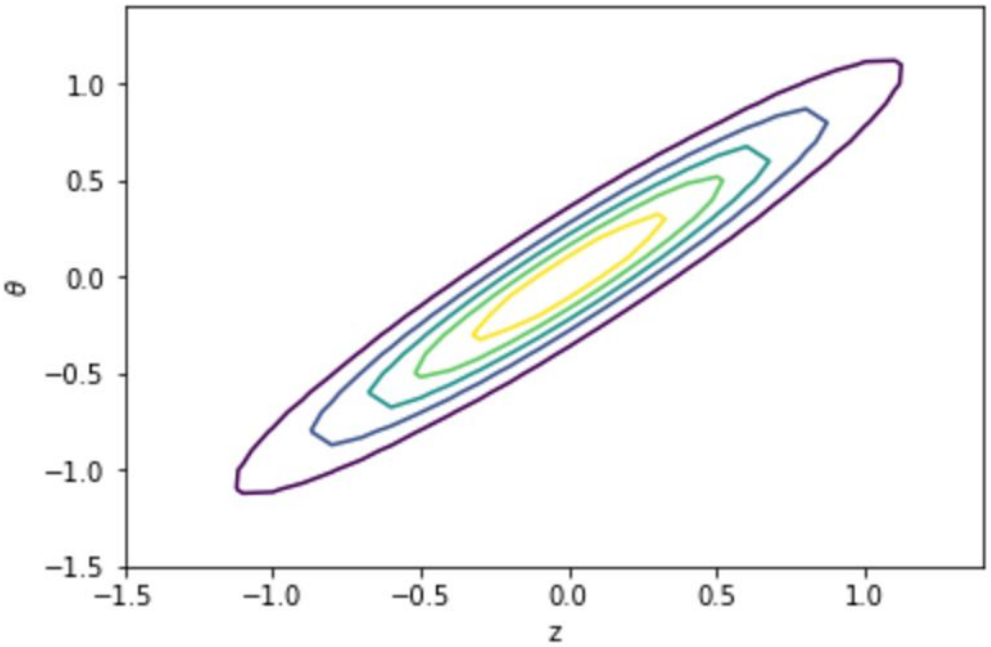
this is a consistent estimate. Note that if we had $\mathrm{Var}_q(\alpha_i) = 0$, then the optimum would have been

$$\begin{aligned}q^{MLE}(\nu) &\leftarrow \mathrm{Gamma}\left(N+1, \frac{S}{4}\right) \\ \mathbb{E}[\nu] &= 4\frac{N+1}{S}\end{aligned}$$

which is too large (i.e., the variance is underestimated), as expected.

Bad bananas

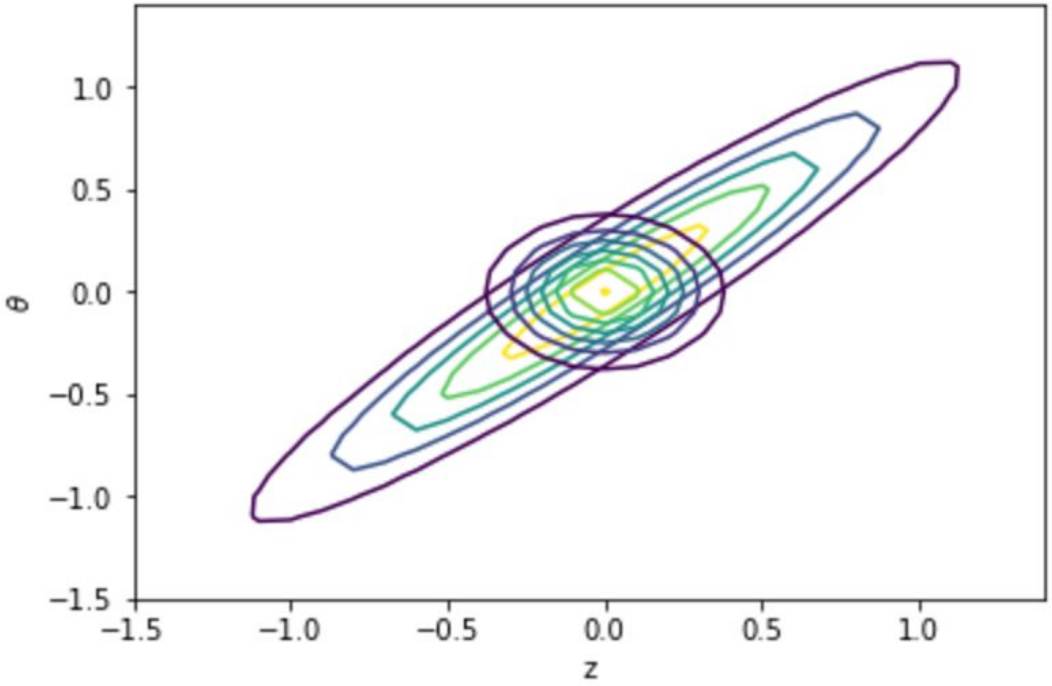
An illustrative example



Let’s start with a correlated Gaussian.

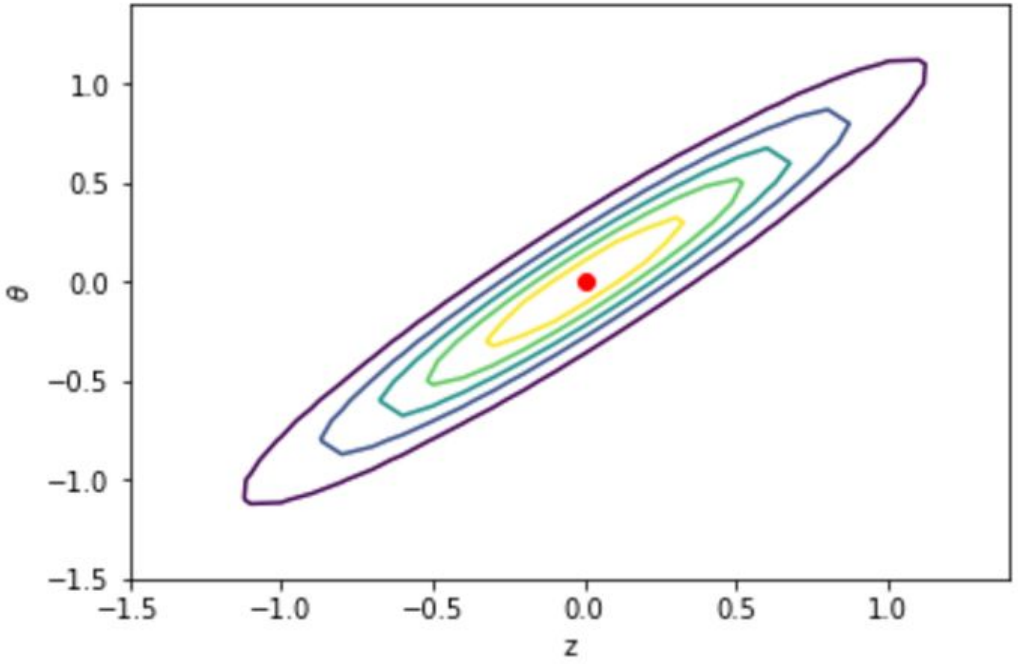
For illustrative purposes we’ll use a one-dimensional latent variable, though of course it’s silly.

An illustrative example



As expected, MFVB gets the means right and the covariances wrong.

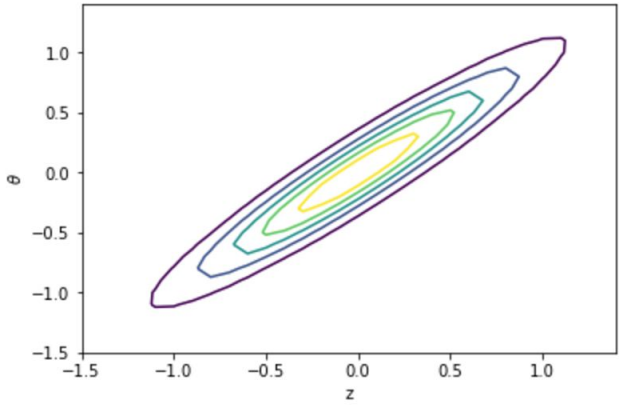
An illustrative example

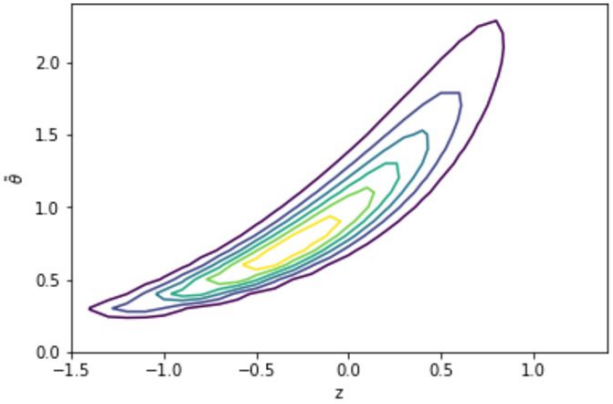


The MAP estimator gets the means right. The negative inverse Hessian will provide a good estimate of the posterior covariance.

This is known as the “Laplace approximation”.

An illustrative example





Suppose we instead had modeled

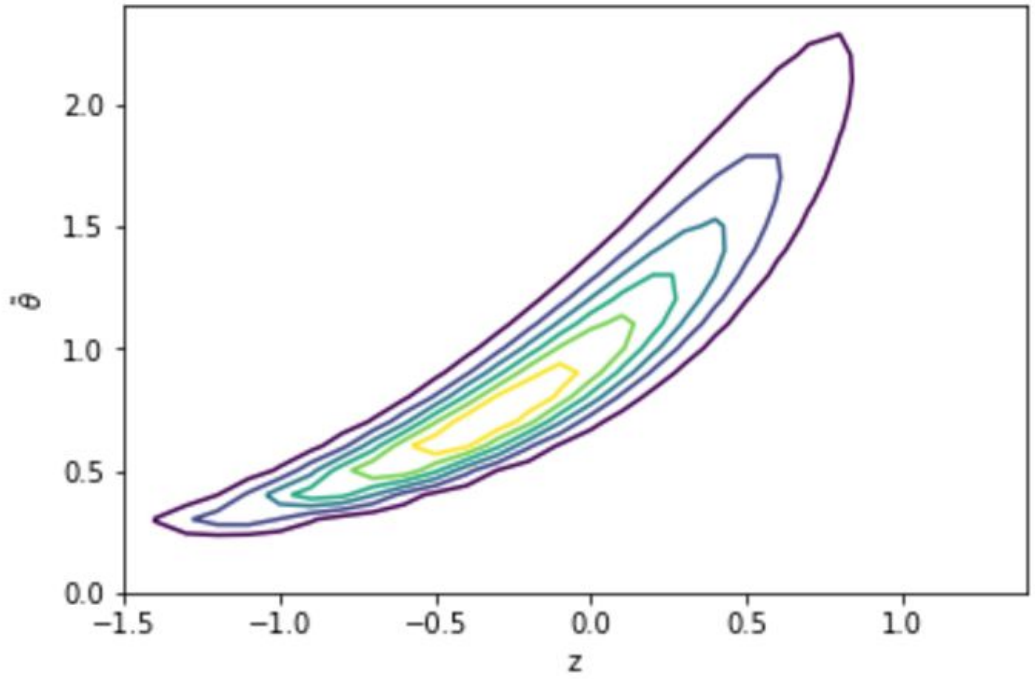
$$\tilde{\theta} = \exp(\theta)$$

$$P_{\tilde{\theta},z}\left(\tilde{\theta}, z\right) = P_{\theta,z}\left(\log\tilde{\theta}, z\right)\frac{d\theta}{d\tilde{\theta}}$$

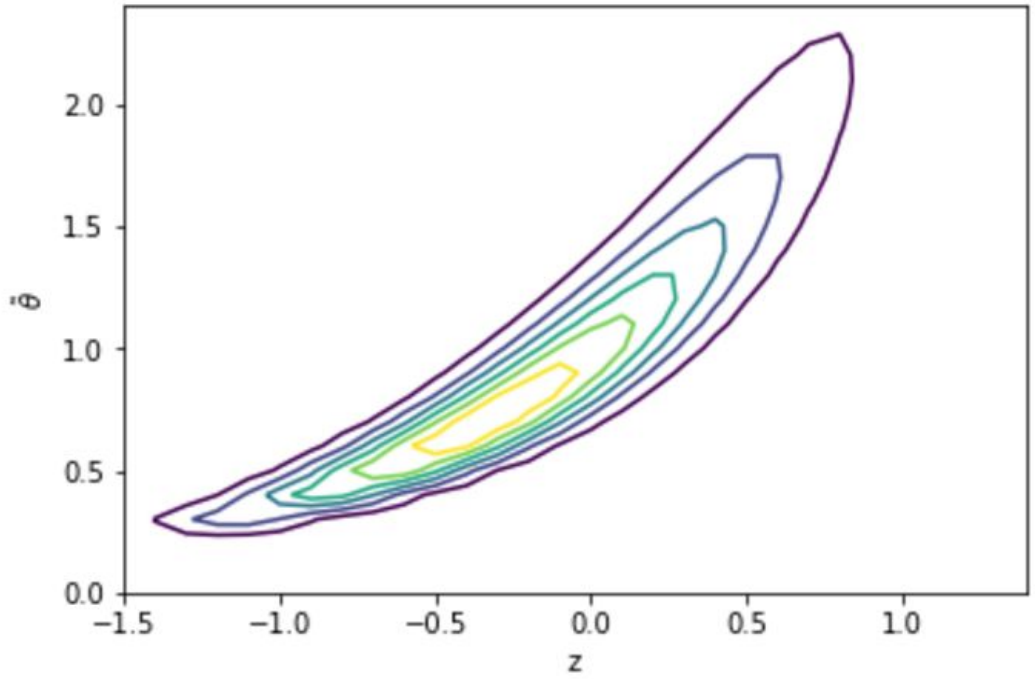
$$= P_{\theta,z}\left(\log\tilde{\theta}, z\right)\exp(-\theta)$$

In $\left(\tilde{\theta}, Z\right)$ space the problem is not as easy.

A bad banana



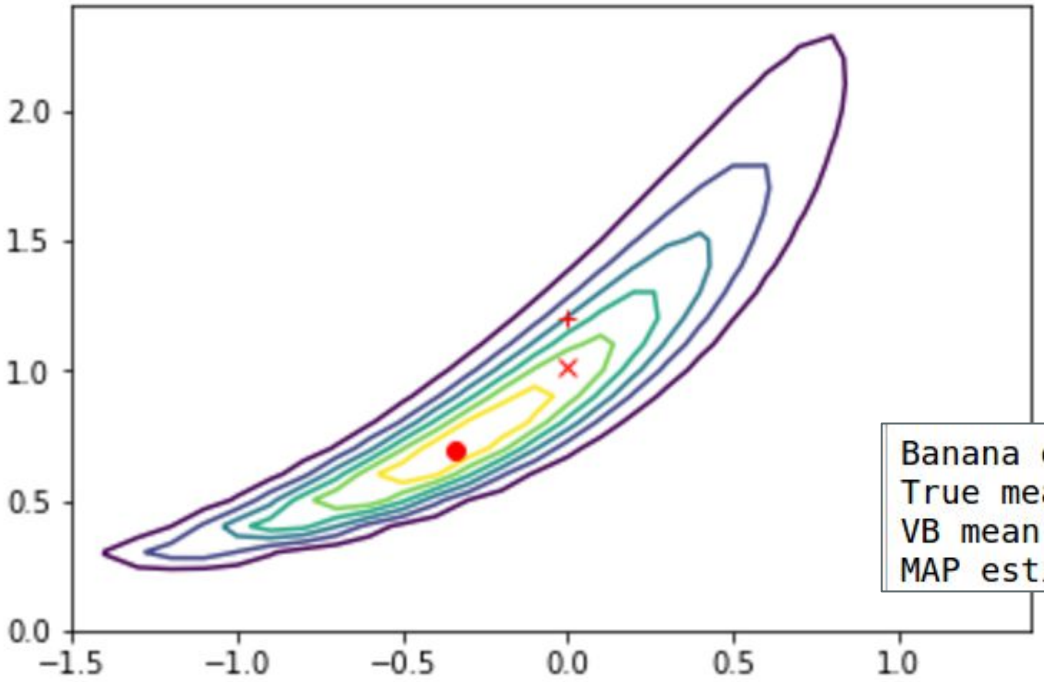
A bad banana



Where do you expect to find:

- The MAP?
- The posterior mean?
- The MFVB estimate?

A bad banana



- o: The MAP
- +: The posterior mean
- x: The MFVB estimate

```
Banana estimates for theta tilde:
True mean:                  1.19721736312
VB mean:                    1.01770490612
MAP estimate:               0.697601425656
```

A bad banana

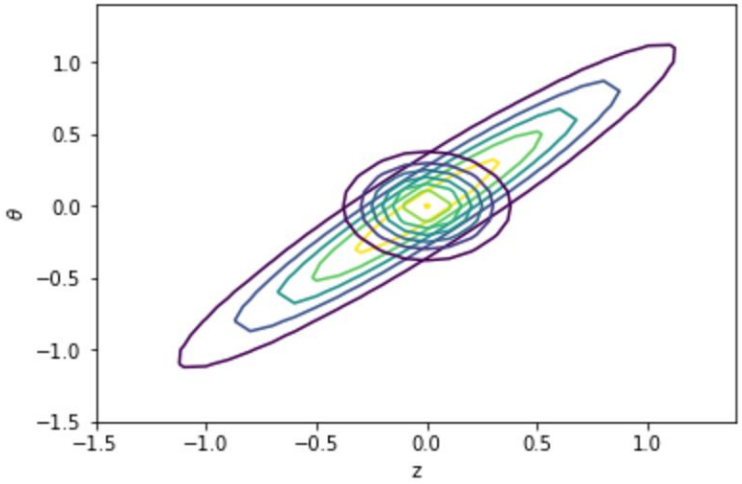
Why did this happen?

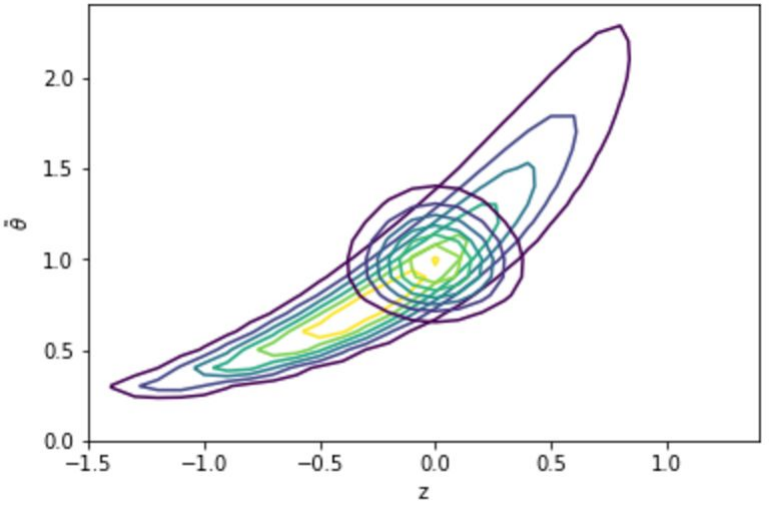
1. The KL divergence is invariant to reparameterization and
2. The mean of a lognormal depends on the variance, which is under-estimated by MFVB.

$$\theta|X \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

$$\exp\left(\theta\right)|X \sim \text{Lognormal}\left(\mu, \sigma^2\right)$$

$$\mathbb{E}\left[\exp\left(\theta\right)|X\right] = \exp\left(\mu + \frac{1}{2}\sigma^2\right)$$



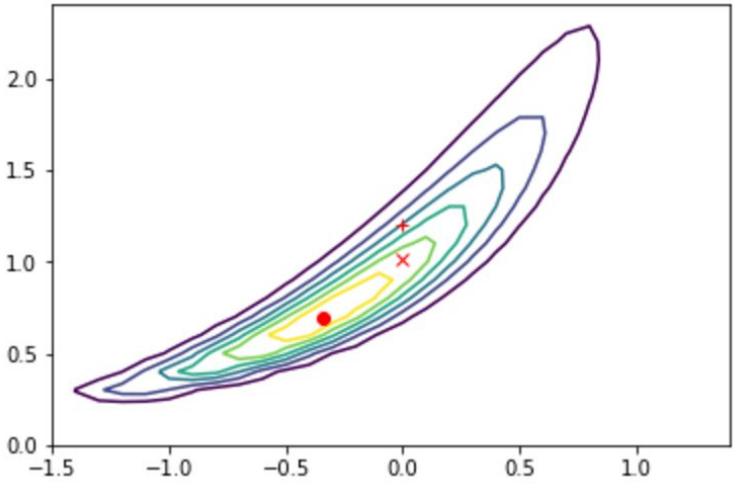


Takeaways

The MAP (or MLE) does not use any information about the variability of the latent variables.

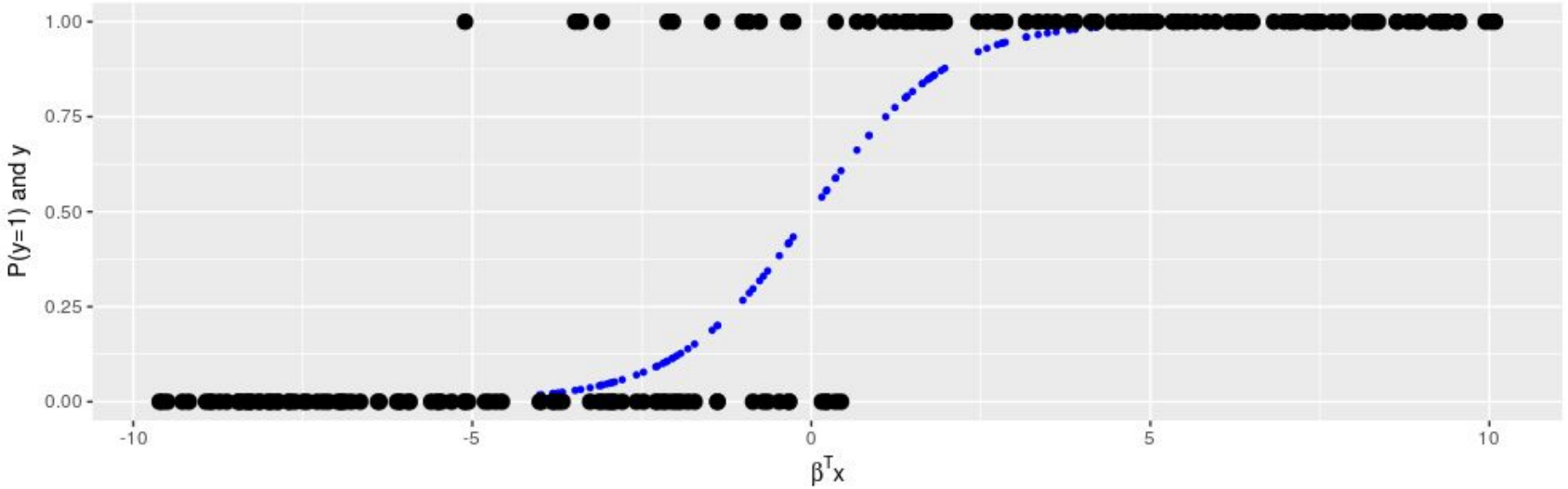
MFVB accounts for some but not all of the variability. It tends to be better than the MAP, but can be worse than a full solution.

The full marginal is the best answer but can be slow.



More realistic example

A logistic generalized linear mixed model for binary outcomes with latent variables:



Local sensitivity, covariances, and variational Bayes, Ryan Giordano, Tamara Broderick, Michael Jordan (in progress)

More realistic example

A logistic generalized linear mixed model for binary outcomes with latent variables:

$$
\begin{aligned}
y_{it}|p_{it},\beta,u_t,x_{it} &\sim \text{Bernoulli}\left(p_{it}\right), \text{ for } t=1,...,T \text{ and } i=1,...,N \\
p_{it} &:= \frac{e^{\rho_{it}}}{1+e^{\rho_{it}}} \\
\rho_{it} &:= x_{it}^T\beta + u_t \\
u_t|\mu,\tau &\sim \mathcal{N}\left(\mu,\tau^{-1}\right).
\end{aligned}
$$

Local sensitivity, covariances, and variational Bayes, Ryan Giordano, Tamara Broderick, Michael Jordan (in progress)

More realistic example

MFVB matches quite well.

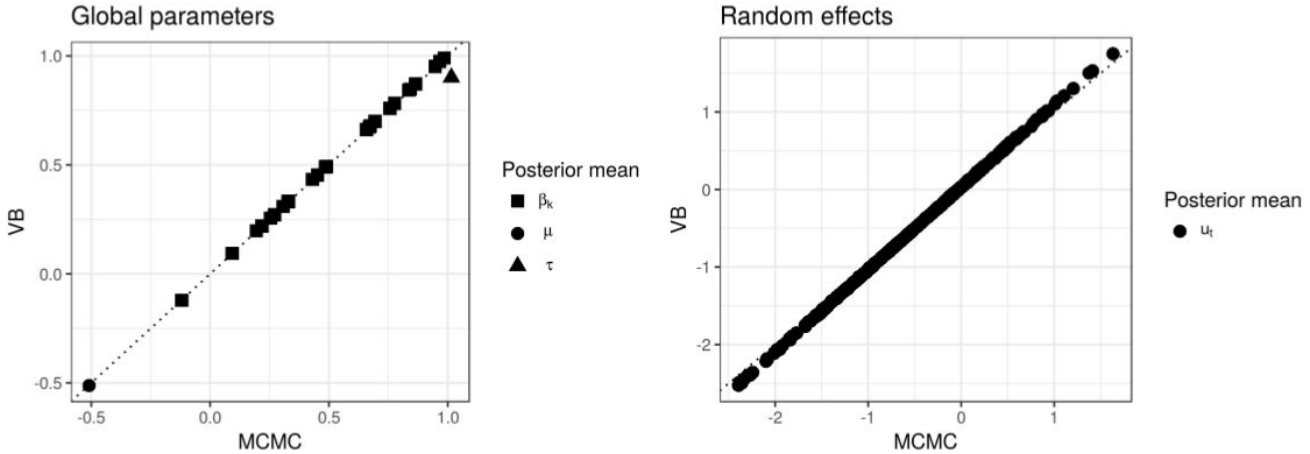
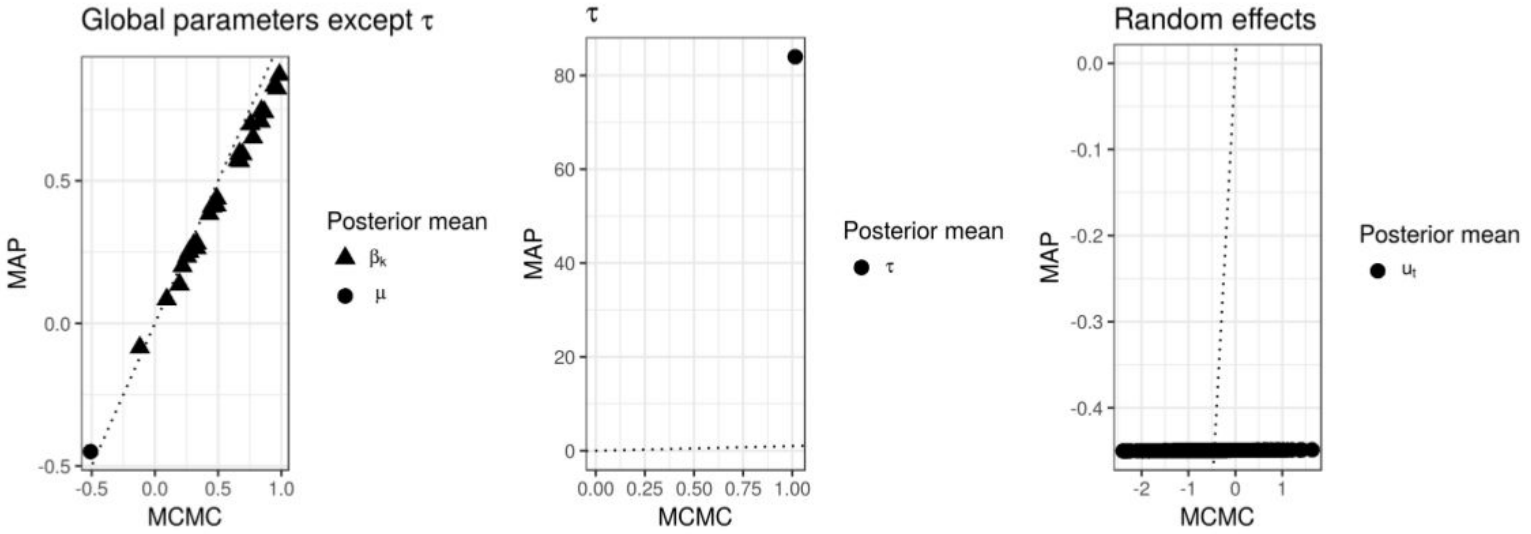


Figure 6: Logit GLMM: Comparison of MCMC and VB Means

Local sensitivity, covariances, and variational Bayes, Ryan Giordano, Tamara Broderick, Michael Jordan (in progress)

More realistic example

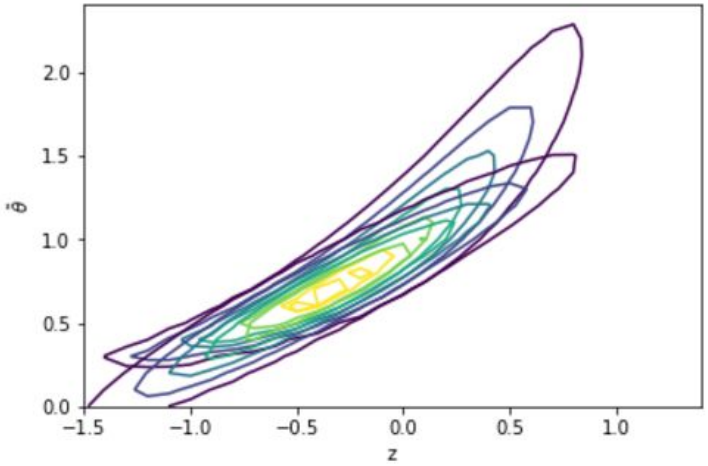
MAP does badly on the latent parameters and their variance.

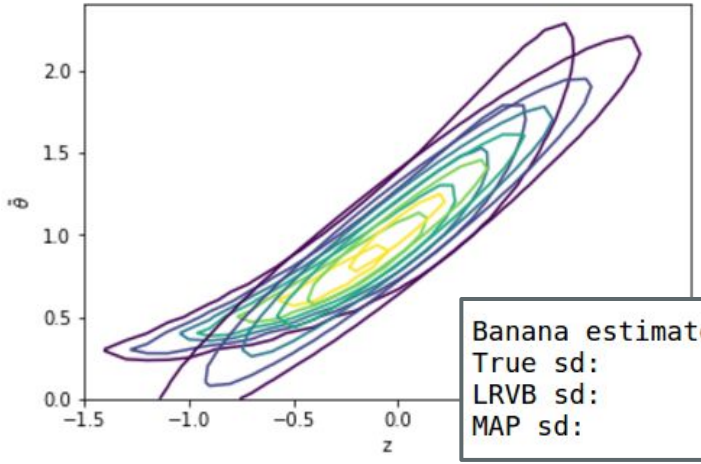


Local sensitivity, covariances, and variational Bayes, Ryan Giordano, Tamara Broderick, Michael Jordan (in progress)

What about covariances?

If the MAP is not meaningful, then the Hessian is not meaningful. But how can you use second-order information with a MFVB solution which is not at a maximum?
See our paper!






```
Banana estimates for theta tilde:
True sd:                    0.788101386932
LRVB sd:                    0.611169325477
MAP sd:                     0.418511159914
```

Linear Response Methods for Accurate Covariance Estimates from Mean Field Variational Bayes: Ryan Giordano, Tamara Broderick, Michael Jordan (2015)

Conclusion

Conclusions

- Sometimes you need to account for the uncertainty in your latent variables, even if you just want point estimates.
- This need and the attendant difficulties are common to both frequentist and Bayesian approaches.
- One can approximately account for the uncertainty in latent variables using variational approaches that minimize KL divergence (VB for Bayes, EM for frequentists).
- These approximations can be better than MAP but still not perfect in the presence of bad bananas.

Questions?